Identification and Functional Characterization of Long Noncoding RNAs Involved in Endosperm Development of *Arabidopsis thaliana*

Quang Trung Do

A thesis submitted for the degree of Doctor of Philosophy

The Department of Molecular and Cellular Biology
School of Biological Sciences

June 2018

Contents

Abstract

Elucidating the molecular events underlying endosperm and embryo development in angiosperms are important both in terms of understanding plant development and developing new methods to enhance crop productivity. Seeds arguably undergo one of the most complex developmental programs of any plant organ, and are therefore subject to many gene regulatory mechanisms. In recent years, it has become clear that various classes of noncoding ribonucleic acid (ncRNA) and covalent histone modifications have important roles in gene regulation. Of these ncRNAs, small RNAs (20 to 25 nucleotides) are beginning to be understood; however, less is known about the role and complexity of long noncoding RNAs (lncRNAs). Here, we detail the methodology for purifying specific cell types, RNA sequencing, bioinformatic annotation of lncRNAs and investigation of biological function, using the reference plant *Arabidopsis thaliana*. We also detail methodology for highly reproducible bisulfite treatment of RNA, efficient locus-specific PCR amplification, detection of 5-methycytosine that includes sequencing on the Illumina MiSeq platform and bioinformatic calling of converted and non-converted cytosines.

Next, we investigated the contribution of transposable elements (TEs) to long intergenic noncoding RNAs (lincRNAs) during plant development and abiotic stress tolerance. Using deep Illumina sequencing, we identified 47, 611 and 398 TE-associated lincRNAs (TE-lincRNAs) from *Arabidopsis*, rice and maize, respectively. We demonstrated that some of these TE-lincRNAs were tissue specifically transcribed and others were expressed after salt, abscisic acid (ABA) or cold treatments. After identification and characterization of about 50 TE-lincRNA mutants, the mutant TE-linc11195 was identified as having less sensitivity to ABA. The TE-linc11195 mutant had longer roots and higher shoot mass when compared with wildtype in the presence of ABA. Our data suggest that TE-lincRNAs might be a promising reservoir to adapt to changing environmental conditions.

We also explored the potential roles of lncRNAs in regulating epigenetic modifications deposited by the Polycomb Repressive Complex 2 (PRC2)

complex. We immunoprecipitated PRC2-associated lncRNAs and sequenced the bound RNAs by Illumina sequencing. We validated the expression of these PRC2-associated lncRNAs by strand-specific reverse transcription polymerase chain reaction (RT-PCR), and computationally predicted their functions in seed development by association with H3K27me3-targeted (PRC2) genes. Interestingly, the data also showed that G-tract motifs (G2L1-4) are significantly enriched among PRC2-binding transcripts. This dataset provides an initial insight into PRC2-associated RNAs and may contribute towards understanding PRC2 function.

Further, we identified 615 lncRNAs from *Arabidopsis thaliana* one day after pollination (DAP) of siliques using high-throughput Illumina sequencing. Next, we showed that some of these lncRNAs could be transcribed in an organ-specific manner or more broadly transcribed in root, flower and silique organs. Among the broadly transcribed lncRNAs, some were differentially abundant, while others were similarly abundant across all three tissue types. We also investigated the function of 42 lncRNAs by using either artificial microRNAs or RNAi to knockdown the targets. Of these, the knockdown plants of *lnc1246* were observed to have smaller cells and organs in all tested tissues: roots, cotyledons and seeds. We also demonstrated with open reading frame analysis that *LNCRNA_1246* was unlikely to encode for a functional protein. Functional analysis using a recessive *lnc1246* mutant allele and reciprocal crosses demonstrated that *LNCRNA_1246* acted maternally to reduce seed size. This could be a result of smaller cells within the outer integument layer and a smaller integument. Together, our results demonstrate that lncRNAs are broadly transcribed and at least one plays an important role in seed size.

Overall, this thesis focuses on the genome-wide identification and characterization of lncRNAs from *A. thaliana* 1DAP silique and the possible functions of lncRNAs in plant development by interacting with their partners, such as TEs and FIS2–PRC2 complexes. It also illustrates the potential effects of lncRNAs on diverse biological processes during plant evolution.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provision of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder (s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search, and also through web search engines, unless permission has been granted by the University of Adelaide to restrict access for a period of time.

Signed…… ……..Date……04/10/2018……..

List of Publications

Trung Do, Zhipeng Qu and Iain Searle (2018) Purification and functional analysis of plant long-non coding RNAs (lncRNA). In press Spinger Science + Business, Media, LLC, New York.

Jun Li, Xingyu Wu, **Trung Do**, Vy Nguyen, Jing Zhao, Pei Qin Ng, Alice Burgess, Rakesh David and Iain Searle (2018) Quantitative and single nucleotide resolution profiling of RNA 5-methylcytosine. In press Spinger Science + Business, Media, LLC, New York.

Wang D, Qu Z, Yang L, Zhang Q, Liu ZH, **Do T**, Adelson DL, Wang ZY, Searle I, Zhu JK (2017) Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in Plants. Plant J. 90:133-146.

Trung Do, Zhipeng Qu, Ashley Jones, Rakesh David, David L. Adelson and Iain Searle (2018) Maternal control of seed size by long noncoding RNA in *Arabidopsis thaliana*. In preparation for The Plant Cell.

Trung Do, Zhipeng Qu, Jun Li, Rakesh David, Chris Heliwell, David L. Adelson and Iain Searle (2018) Identification of PRC2-associated long noncoding RNA in *Arabidopsis thaliana*. In preparation for BMC Plant Biology.

Abbreviations

A

ANRIL	Antisense noncoding RNA in the INK4 locus
AIR	Antisense Igf2r RNA
ASL	Antisense long
APOLO	Auxin-regulated promoter loop RNA
AGO4	Protein Argonaute 4
AG	Agamous
AGL62	*Arabidopsis* thaliana MADS-box protein AGL62
ASCO-lncRNA	Alternative splicing competitor long noncoding RNA

C

CC	Central cell
ChIP	Chromatin immunoprecipitation
COOLAIR	Cold-induced long antisense intragenic RNA
COLDAIR	Cold-assisted intronic noncoding RNA
CMT3	Chromomethylase 3
COLDWRAP	Cold of winter–induced noncoding RNA from the promoter
CHIRP	Chromatin isolation by RNA purification
CLF	Curly leaf

D

DME	DNA glycosylase DEMETER
DAP	Day after pollination

E

EC	Egg cell
ENOD40	Early nodulin 40
ELENA 1	ELF18-induced long noncoding RNA 1
ESRP2	Epithelial splicing regulatory protein 2
EDE1	Endosperm defective 1
easiRNA	Epigenetically activated small interfering RNA
EMF	Embryonic flower

EZH2	Enhancer of zeste homolog 2
EED	Embryonic ectoderm development

F

FENDRR	FOXF1 adjacent noncoding developmental regulatory RNA
FIRRE	Firre intergenic repeating RNA element
FLD	Flowering locus D
FIE	Fertilization independent endosperm
FWA	Flowering wageningen
FLC	Flowering locus C
FIS2	Fertilization independent seed 2

H

HDAC3	Histone deacetylase 3
HOTAIR	HoxA transcript antisense RNA
HOTTIP	HoxA transcript at the distal tip
hnRNP-K	Heterogeneous nuclear ribonucleoprotein K

I

INTACT	Isolation of nuclei tagged in specific cell types

J

JARID2	Jumonji/AT-rich interactive domain 2

L

LHP1	Like heterochromatin protein 1
LncRNA	Long noncoding RNA
LSD1	Lysine-specific demethylase1
LEC2	Leaf cotyledon2
LincRNA	Long intergenic noncoding RNA
LSD1	Lesion simulating disease 1
LNCRNA_1246	LNCRNA_1246 genes
lnc1246	Mutant of *LNCRNA_1246*

M

MEG	Maternally expressed genes
miRNA	Micro RNA
MLL-WDR5	Mixed lineage leukaemia-WD-repeat protein-5
MED19a	Mediator of RNA polymerase II transcription subunit 19a-like protein
MLE	Maleless
MSL	Male-specific lethal
MEA	Medea
MALAT1	Metastasis-associated lung adenocarcinoma transcript-1
MET1	Methyltransferase 1
MSI1	Multicopy suppressor of IRA 1

N

NSR	Nisin resistance protein
NEAT1	Nuclear paraspeckle assembly transcript 1
ncRNA	Noncoding RNA

P

PEG	Paternally expressed genes
PcG	Polycomb group protein
PRC1	Polycomb recessive complex 1
PHE1	Pheres1
Pol II	Polymerase II
Pol IV	Polymerase IV
PRC2	Polycomb recessive complex 2
PREs	Polycomb response elements
PSPC1	Paraspeckle component 1
PSF	Polypyrimidine tract-binding protein-associated splicing factor
PURA	Purine-rich element-binding protein
PANDA	P21-associated ncRNA DNA damage activated

R

RBP	RNA binding protein
RBD	RNA binding domain
RIP	RNA immunoprecipitation
Rox1	RNA on the X1
TERC	Telomerase RNA component

S

SRSF1	Serine/arginine-rich splicing factor 1
snoRNA	Small nucleolar RNA
siRNA	Small interfering RNA
SHARP	SMRT- and HDAC-associated repressor protein
STM	Shoot meristemless
SUZ12	suppressor of zeste 12

T

TCAB1	Telomerase and Cajal body protein 1
TEs	Transposable elements
tRNA	Transfer RNA
tasi-RNA	Trans-acting siRNA

V

VRN	Vernalisation

W

WUS	Wuschel

X

Xist	X-inactive specific transcript

Acknowledgements

Firstly, a special thank you to my PhD supervisor, Iain Searle. Thank you for your discussion, feedback, guidance and encouragement over the years—they really mean a lot to me. I have learnt a lot from you. Thank you.

I would also like to thank my co-supervisor, Rakesh David, who provided invaluable feedback and guidance. This sustained me during the tough times, helping me to learn and achieve. I have really enjoyed working with you, and you have taught me a lot. Thank you.

I would like to thank other members of the Searle laboratory, past and present, for making the laboratory such a supportive and enjoyable place to work. Moreover, I would like to thank our collaborator, Zhipeng Qu, for undertaking bioinformatics analysis and generating beautiful figures.

I would like to thank the Elite Editing™ for proof-reading and editing of my research Thesis (Preface section, Chapter 5 and Chapter 6).

I would like to thank my best friends, Phuong Nguyen, Tan Dang and Vy Nguyen, for encouraging and supporting me throughout my PhD. There are many others I am lucky to have that helped me progress through difficult times.

Thank you to my loving family for always being there, and for encouraging and supporting me throughout my PhD. Thank you for listening to my whining and for cheering me up when experiments were failing epically. Your love and support throughout my whole life means everything to me. I would like to use my thesis as a gift to my children, Quang Long and Minh Chau.

Finally, I would like to thank and recognise my sponsors, the University of Adelaide and Vietnam International Education Cooperation Department, who provided me with the postgraduate scholarship to support me during my PhD.

Chapter 1: Introduction

In modern Western societies, seeds such as cereal grains, oilseeds and legumes serve as important sources of carbohydrates, lipids and proteins (Venglat et al., 2014). Reliance on a limited number of crops could lead to problems with global food production and security because of over-population, climate change and other adverse factors (Beddington, 2010). Therefore, we need more progressive improvements in both breeding methods and ways of exploiting crop germplasm resources to produce new related traits with special characteristics such as increased crop productivity, bigger seed size, higher nutrition quality or better resistance to environmental stressors. Because important factors controlling seed traits relate to gene expression and regulation, this project concentrates on the epigenetic mechanisms that regulate seed size in *Arabidopsis thaliana* through the regulation of endosperm development.

1.1 Introduction to the Model Plant *Arabidopsis thaliana*

Arabidopsis thaliana is a small, self-fertilizing plant of the Brassicaceae that requires simple growth conditions and produces thousands of seeds in a short generation time of six weeks; significantly, its small genome has been fully sequenced (Somerville and Koornneef, 2002). In addition, *Arabidopsis* ecotypes vary in many morphological and physiological traits that provide a useful resource for identifying the molecular basis of complex traits by exploiting the polymorphisms in nucleotide sequences and epigenetic variation. Through The *Arabidopsis* Information Resource (TAIR) website, *Arabidopsis* researchers can obtain information about protein-coding and noncoding genes, markers, clones and nucleotide polymorphisms, and can access DNA and seed stocks (Garcia-Hernandez and Reiser, 2002). In this project, I used *Arabidopsis* as a model representative flowering plan to study seed development.

1.2 *Arabidopsis thaliana* Seed Development

Earth's land is mostly covered by plants, of which three-quarters are flowering plants (angiosperms). Over the course of evolution, angiosperms have developed

a wonderful reproductive strategy in which the embryo is protected and supplied with nutrients during germination. A typical seed structure includes three main parts: the embryo, endosperm and seed coat. Although there is a significant difference in the storage component between seeds of monocots and dicots—which have an endosperm and embryo respectively—the seed developmental program, which includes fertilization and embryo and endosperm development, is conserved (Mosher and Melnyk, 2010; Venglat et al., 2014).

Seed development involves a complicated interplay between the embryo, endosperm and seed coat that is activated by double fertilization. Double fertilization is a process in which one pollen sperm cell fertilizes a haploid egg, forming a diploid embryo and the other sperm fertilizes a homodiploid central cell of the ovule, forming a triploid endosperm (Hamamura et al., 2012). The embryo and endosperm are protected by inner and outer integuments of the ovule (Fig. 1 below). As a result of double fertilization, genome dosage in the early stage of seed development differs.

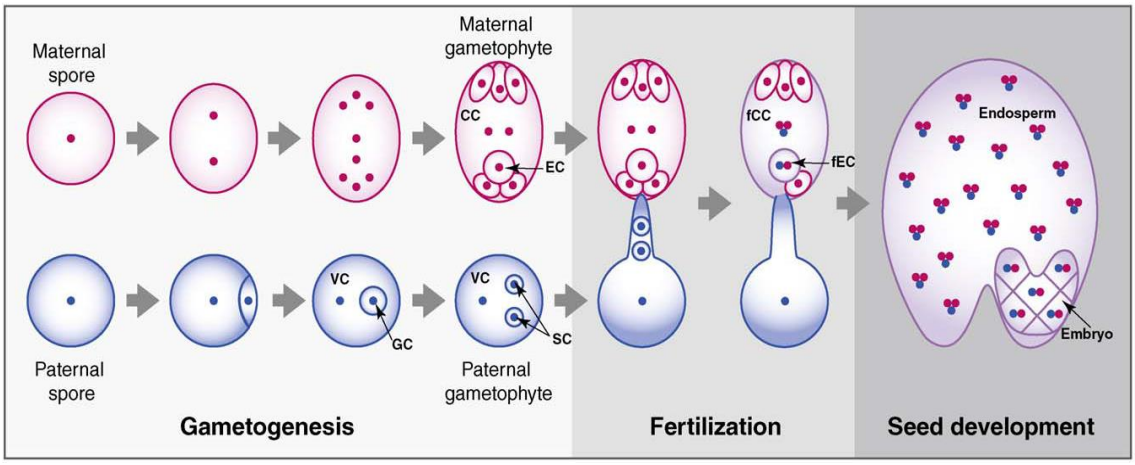


Figure 1. Overview of double fertilization. During gametogenesis, the maternal spore undergoes several mitotic rounds to form the haploid egg cell (EC) and the homodiploid central cell (CC) in the ovule; while two sperm cells (SC) are formed in the paternal spore from the generative cell (GC), which is engulfed by the vegetative cell (VC). Fertilization is initiated when the growing pollen tube bursts near the ovule, where one sperm cell fertilizes an egg cell to form fertilized egg cell (fEC) developing into the diploid embryo, and the other sperm cell fertilizes the central cell to form the fertilized central cell (fCC) growing to a triploid endosperm (Mosher and Melnyk, 2010).

Double fertilization is followed by a morphogenetic phase during which the zygote and endosperm are genetically programmed to form the embryonic body plan

and nutritive tissue, respectively (Fig. 1). This is followed by a maturation phase during which the seed accumulates nutrients such as carbohydrates, lipids, proteins and several important nutrients including vitamins and minerals (Fig. 2). The seed finally desiccates and enters the dormancy phase of angiosperms (Jenik et al., 2007; Sreenivasulu and Wobus, 2013).

1.2.1 Embryo development in *Arabidopsis thaliana*

Fusion of haploid sperm and egg cells produces a diploid zygote that then undergoes cell division and differentiation to produce a suspensor (cylindrical structure) during the pre-global embryo stage. The terminal suspensor cell undergoes further cell divisions and differentiation to produce a global and then a heart stage embryo. At this stage the plant body plan is defined and consists of cotyledons, hypocotyl and root and shoot meristems (see Fig. 2 and Table 1 for a summary).

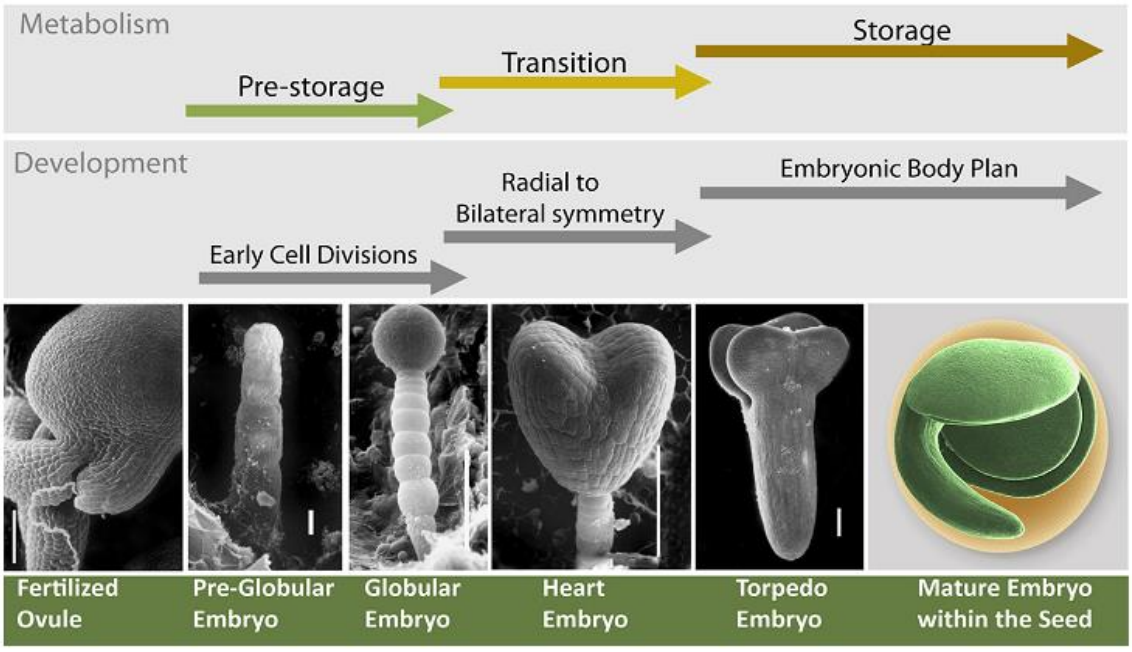


Figure 2. Overview of the major stages of embryo development in *Brassica.* Inside the ovule integuments, initially endosperm development occurs to support the developing embryo; in parallel the zygote divides to form the suspensor, then the global embryo and later the heart stage embryo that defines the embryonic body plan. Metabolic programs to synthesize and later store the necessary nutrients for seed maturation are expressed (Venglat et al., 2014).

Genetic screens in *A. thaliana* have identified genes required for embryogenesis, which include *LEAFY COTYLEDON2* (*LEC2*) a transcription factor required to

induce embryo development; *GNOM*, which encodes a guanine nucleotide exchange factor mediating subunit interaction with cyclophilin 5; *SHOOT MERISTEMLESS* (*STM*), which is required for shoot apical meristem function; *MONOPTEROS,* a transcription factor that plays a mediator role in embryo formation and vascular development; and *FACKEL,* which functions in cell division and expansion (Allan and Abed, 2002). These genetic studies have laid the foundation of a genetic framework for embryogenesis. Recent cell-specific transcriptional profiling (Palovaara et al., 2018) has elegantly increased our knowledge of the spatial transcriptional networks compared with previous datasets (Harada et al., 2010; Radoeva et al., 2016).

Interestingly, early embryo development requires a signal from the endosperm. Using reverse genetics and biochemical approaches it has been demonstrated that a peptide signal, EMBRYO SURROUNDING FACTOR 1 (ESF1), accumulates in the central cell and embryo-surrounding endosperm cells to act in a non-cell autonomous manner to promote suspensor elongation in the YODA mitogene-activated protein kinase pathway (Costa et al., 2014). Later embryo development is also strongly influenced by endosperm-derived nutrients and the exchange of signal molecules between endosperm and embryo (see Section 1.3).

1.2.2 Endosperm development in *Arabidopsis thaliana*

The endosperm is one of three components of a typical seed and plays a critical role in seed development, where it functions to nourish the embryo. According to Berger (1999), endosperm development can be divided into four phases: syncytial, cellularization, differentiation and programmed cell death. However, the duration of each phase differs between species and there is overlap between each phase (Berger, 1999).

In *Arabidopsis*, there are two distinct phases during endosperm development: the syncytial phase and the cellularized phase (Berger, 2003; Li and Berger, 2012). During the syncytial stage, the triploid zygotic nucleus successfully carries out hundreds of mitotic divisions without cytokinesis, producing a large cell containing many hundreds of nuclei. The latter stage is the cellularization stage, which is

initiated at the heart stage of embryogenesis and is characterized by a cell wall being formed between two close nuclei; cellularization occurs from the micropylar endosperm to peripheral endosperm, but not in the chalazal endosperm, which remains in the syncytial endosperm until seed maturation (Boisnard-Lorig et al., 2001; Costa et al., 2004) (see Figure 3 and Table 1 for summary). Notably, the molecular trigger for endosperm cellularization is unknown.

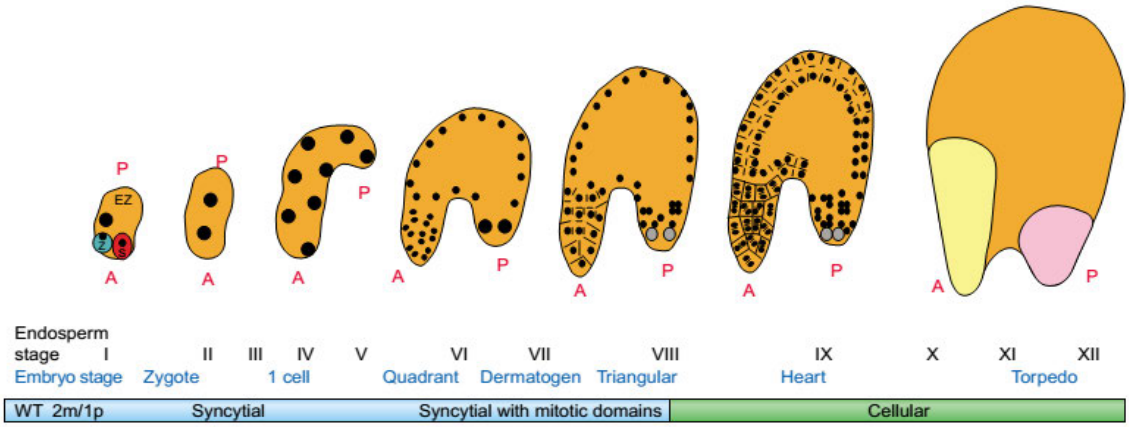


Figure 3. Endosperm development in *Arabidopsis*. The first stage is syncytium (blue); the cell undergoes mitosis without cytokinesis, resulting in a large cell with many hundreds of nuclei. The second stage, cellularization (green), includes the cytokinesis events (A–anterior pole; P–posterior pole; Z- zygote; EZ-endosperm zygote; S-synergid; yellow domain: micropylar endosperm; orange domain: peripheral endosperm; pink domain: chalazal endosperm) (Berger, 2003).

Table 1. Summary of endosperm development in *Arabidopsis* (Boisnard-Lorig et al., 2001)

Endosperm Developmental Stage	No. of Endosperm Nuclei	Embryo Developmental Stage	Time after Fertilization (hr)	Cytological Events
I	1	Zygote	0	One large nucleus close to the zygote.
II	2	Zygote	2–4	First syncytial mitosis. The two nuclei migrate to opposite poles of the endosperm.
III	4	Zygote	ND[a]	Second syncytial mitosis. Nuclei division planes are parallel to the main polar axis.
IV	(6)–8	Elongated zygote	ND	Third syncytial mitosis. Nuclei division planes are perpendicular to the main polar axis. One or two nuclei migrate to the chalazal pole, and the six or seven other nuclei become tightly linked to the endosperm peripheral cell wall. Nuclei of the CZE do not enter mitosis together with nuclei of the PEN.
V	12–16	Elongated zygote, one cell	ND	Fourth syncytial mitosis. Nuclei divisions are not synchronous, and nuclei of the CZE are larger than nuclei in the PEN.
VI	24–28	One cell, two cell	12–18	Fifth syncytial mitosis. The CZE contains one to four large nuclei. Nuclei in the MCE undergo synchronous divisions earlier than nuclei in the PEN.
VII	44–48	Two cell, quadrant, octant	24	Sixth syncytial mitosis. The delay between nuclei division in the MCE and the PEN becomes more pronounced. In the PEN, coordinated nuclei divisions take place as a wave. The PEN and the MCE become independent domains of cyclin B1;1 expression.
VIII	90	Octant	30	Seventh syncytial mitosis. A layer of nuclear cytoplasmic domains surrounds the embryo, and the MCE is the only part of the endosperm with two layers of nuclear cytoplasmic domains.
IX	200	Dermatogen-globular	36–60	Eighth syncytial mitosis restricted to the PEN. Divisions in the MCE are completely independent of divisions in the PEN. The CZE contains large and small nuclei.

[a] ND, not determined.

1.3 Embryo Development is Influenced by Endosperm Developmental Transitions

It is clear that after fertilization, the synchronous division of the endosperm nucleus in the syncytial phase, together with integument cell proliferation and elongation, leads to a rapid increase in seed size (Li and Berger, 2012). After the syncytial phase, endosperm cellularization occurs and is followed by embryo growth utilizing the nutrients from the endosperm. The growing embryo invades the former endosperm volume (Fig. 4a&4b).

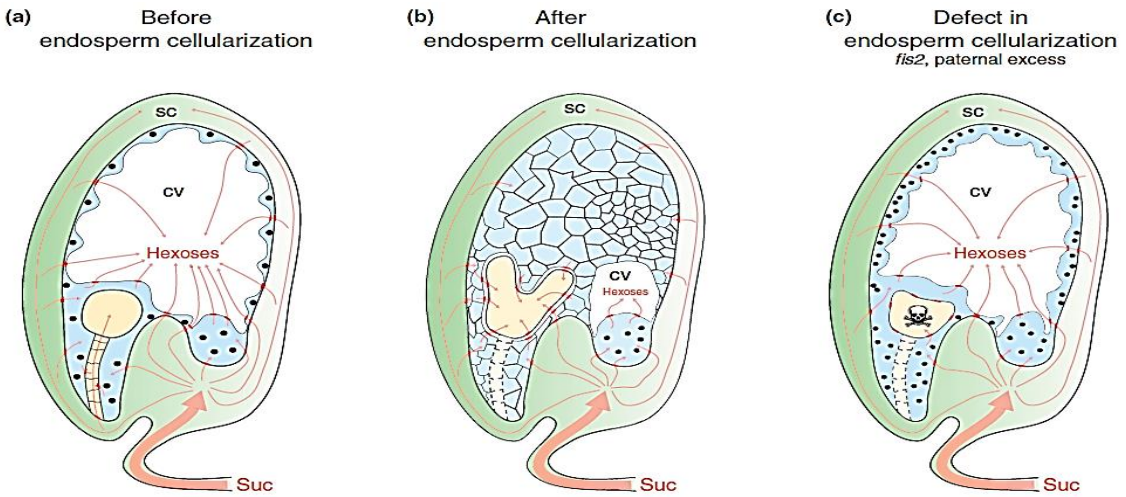


Figure 4. Embryo survival is dependent on endosperm cellularization. Nutrients (sucrose in this case) are transferred from the mother to the embryo throug the integument and endosperm (SC: seed coat, CV: central vacuole, Suc: sucrose, black circles: nuclei, red bars: sucrose transporters). (a) In young seeds, the vacuole makes up the largest proportion of the endosperm, which is surrounded by a thin layer of the syncytium cytoplasm. Sucrose is transferred quickly into the vacuole through the integument and the thin syncytium cytoplasm. The embryo may obtain sucrose via the suspensor or surrounding endosperm via the suspensor. (b) At a later stage, cellularization causes the vacuole to shrink and decreases sucrose transport to the vacuole. Sucrose is transferred directly from the endosperm to the embryo through the sucrose transporters, which are expressed on the cell of the embryo-surrounding region and the embryo epidermis. The suspensor is degraded. (c) In case of a defect in endosperm cellularization, the endosperm is still occupied by the central vacuole at later stages of seed development. Consequently, the sucrose is maintained in the central vacuole but the sucrose supply for the embryo is reduced, causing reduced embryo growth and death (Lafon-Placette and Köhler, 2014).

Endosperm cellularization has been shown to be important for embryo viability. For example, mutation in endosperm-specific *FERTILIZATION-INDEPENDENT SEED 2 - POLYCOMB RECESSIVE COMPLEX 2 (FIS2-PRC2)* leads to aborted seeds that fail to undergo endosperm cellularization and contain embryos arrested at the heart stage of development (Fig. 4c) (Chaudhury et al., 1997). Embryo development is also affected by perturbations in endosperm development. For example, the *endosperm defective 1 (ede1)* mutant causes failure of endosperm cellularization, leading to defects in embryo and overall seed development (Hehenberger et al., 2012). An extensive list of mutations affecting endosperm and seed development is provided in Appendices. Moreover, endosperm cellularization shrinks the large central vacuole, which is the major storage compartment for hexoses in the seed and determines sink strength

during early seed development by rapidly converting imported sucrose into hexoses, likely mediated by activity of vacuole-localized invertases (Morley-Smith et al., 2008; Hehenberger et al., 2012; Lafon-Placette and Köhler, 2014). Hence, endosperm cellularization will cause a reduction of sink strength of the central vacuole, which might be a signal allowing the embryo to establish itself as the major sink in the seed. This is consistent with studies that demonstrated that rapid embryo growth and storage product accumulation starts only after endosperm cellularization (Morley-Smith et al., 2008; Baud et al., 2008). Therefore, failure of endosperm cellularization might result in an undersupply of sucrose for the embryo, as sucrose remains to be transported to the central vacuole. This implies that the timing of endosperm cellularization plays an important role in determining the final seed size as endosperm cell divisions cease strictly before cellularization. In addition, *agl62, fis2* and *fie* mutants affect the timing of cellularization, suggesting that these genes play important roles during endosperm development (Vinkenoog et al., 2003; Hehenberger et al., 2012).

Overall, the timing of endosperm cellularization plays an important role in embryo viability, which can be explained by a sink–source relationship as well as the exchange of signal molecules between endosperm and embryo.

1.4 Developmental Timing of Endosperm Development is Partially Controlled by the Polycomb Group Complex

Endosperm celllularization is a process that appears at the end of the syncytial phase when the nucleus has carried out eight mitotic division, and is characterized by the formation of cell walls among nuclei forming individual cells. This phenomenon is triggered from the anterior to peripheral domains and does not appear in chalazal endosperm, which is thought to have a role in transferring maternal nutrients to the embryo (Costa et al., 2004; Li and Berger, 2012). The signal to activate endosperm cellularization is proposed by the critical nucleo-cytoplasmic ratio, based on results from mutants that have fewer cells in the endosperm as a result of early cellularization (Li and Berger, 2012).

Interestingly, in the *fis2* mutant, two enzymes—pectinesterase and glycosyl hydrolase—were found to be deregulated, which in wild-type degrades the major

components of plant cell walls; this might be the underlying cause of endosperm cellularization failure in *fis* mutants (Weinhofer et al., 2010) (Figure 5).

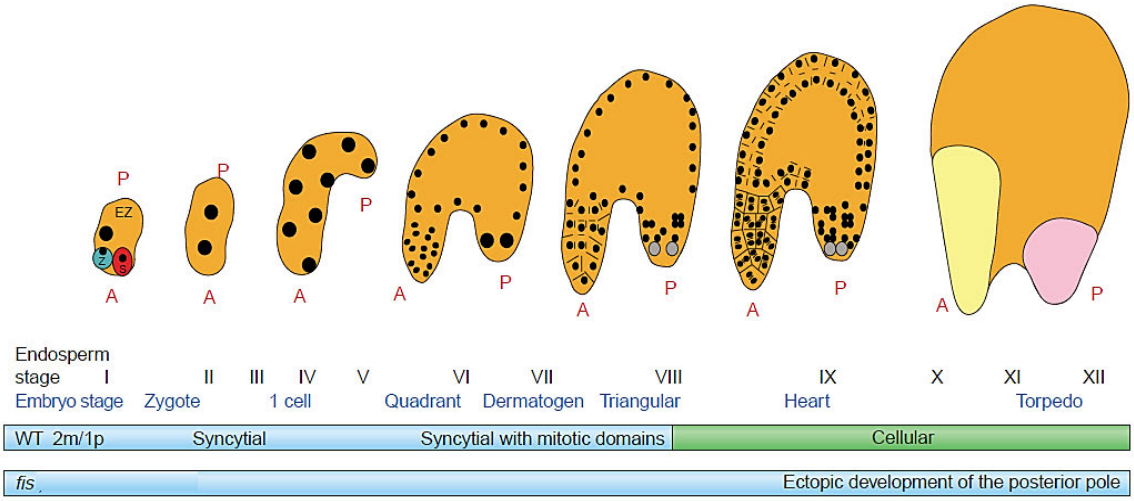


Figure 5. Endosperm development in a *fis* mutant. The upper bar indicates normal endosperm development, which includes two phases: syncytium (blue) and cellularization (green). The lower bar indicates endosperm development in the *fis* mutant, which exhibits only the syncytium phase. A–anterior pole; P–posterior pole; Z- zygote; EZ-endosperm zygote; S-synergid; Yellow domain: micropylar endosperm; Orange domain: peripheral endosperm; Pink domain: chalazal endosperm (Berger, 2003).

The timing of endosperm cellularization has been shown to correlate with the extension of nuclear proliferation and may affect seed size, sink strength and grain weight (Kang et al., 2008; Lafon-Placette and Köhler, 2014; Orozco-Arroyo et al., 2015). Genetic mutants impaired in endosperm cellularization exhibit different effects in their embryo and seed development. For example, the *knolle, hinkel, open house, runkel* and *pleiade* mutants affect cytokinesis in the embryo (Sorensen et al., 2002), while the *spätzle* and *ede1* mutants delay embryo development at the heart stage, leading to seed abortion (Sorensen et al., 2002; Pignocchi et al., 2009). This shows that endosperm cellularization plays a crucial role in embryo and seed development. Recently, based on analysis of those mutations, three redundant pathways regulating endosperm cellularization were proposed (Kang et al., 2013; Orozco-Arroyo et al., 2015). The first pathway is based on the action of *APETALA2* and the MADS-box transcription factor AGL62 (Kang et al., 2008). The second pathway includes members of the polycomb group (PcG) proteins and their targets. The third pathway is the IKU pathway, which involves the activities of several genes including *HAIKU1 (IKU1), HAIKU2*

(IKU2), SHORT HYPOCOTYL UNDER BLUE1 (SHB1) and *MINISEED3 (MINI3)*. These independent networks act as key regulators of endosperm development by regulating the timing of endosperm cellularization, with a major effect on final seed size. Moreover, the timing of endosperm cellularization is also affected by interploidy crosses (increased maternal or paternal genome doses)—which can deregulate cellularization—and is thought that the PcG pathways influence maternal excess (Hehenberger et al., 2012), while the Polymerase IV (Pol IV)-dependent epigenetically activated small interfering RNA (easiRNA) pathways influence paternal excess (Borges et al., 2018; Martinez et al., 2018).

PcG proteins are a family of proteins responsible for cellular differentiation during development via transcriptional repression (Farrona et al., 2008). PcG protein complexes are conserved in plants and animals (Farrona et al., 2008). PcG proteins have two important complexes—polycomb repressive complex 1 (PRC1) and polycomb repressive complex 2 (PRC2)—that function sequentially to repress target genes. PRC2 modifies the chromatin by tri-methylating the lysine amino acid residue located at position 27 of the amino-terminal tail of histone H3 (Simon and Kingston, 2009; Schuettengruber and Cavalli, 2009). The resulting repressive histone 3 lysine 27 trimethylation (H3K27me3) modification acts as a label to recruit PRC1 (Schuettengruber and Cavalli, 2009). In *Arabidopsis*, the diverse PRC2 subunit homologues probably form at least three different PRC2-like complexes with distinct functions: (1) the EMBRYONIC FLOWER (EMF) complex includes CURLY LEAF/SWINGER (CLF/SWN), EMBRYONIC FLOWER 2 (EMF2), FERTILIZATION-INDEPENDENT ENDOSPERM (FIE) and MULTICOPY SUPRESSOR OF IRA 1 (MSI1), which have roles in promoting vegetative development of the plant and delaying reproduction, as well as maintaining cells in a differentiated state (Yoshida et al., 2001; Chanvivattana et al., 2004); (2) the VERNALIZATION (VRN) complex consists of CLF/SWN, VRN2, FIE and MSI1. VRN has functions in establishing epigenetic silencing after vernalization, and enables flowering (Chanvivattana et al., 2004; De Lucia et al., 2008); (3) the FIS complex includes MEDEA (MEA), FIS2, FIE and MSI1, and has been shown to have functions in preventing seed development in the absence of fertilization, and is required for normal seed development (Köhler et al., 2003; Weinhofer et al., 2010). Moreover, the FIS2–PRC2 complex has been

shown to be a key regulator of endosperm development by regulating the timing of endosperm cellularization, with a major effect on final seed size (Köhler et al., 2003; Weinhofer et al., 2010).

Therefore, studying mechanisms of the PcG pathway that influence the timing of endosperm cellularization is very important for understanding underlying mechanisms that control seed development as well as seed size.

1.5 Molecular Mechanisms Controlling Endosperm Gene Expression

Genomic imprinting in mammals and flowering plants is an epigenetic phenomenon leading to allele-specific expression depending on the parent of origin (Vinkenoog et al., 2003; Feil and Berger, 2007). It has an essential role in normal growth and development and has potentially evolved as a mechanism to balance parental resource allocation to the offspring (Haig and Westoby, 1989). The maternally expressed imprinted genes (MEG) are suggested to reduce nutrient flow to the embryo whereas the paternally expressed imprinted genes (PEG) promote nutrient flow to the embryo (Haig and Westoby, 1989). Early in gametogenesis, the alleles of imprinted genes are differentially modified with one or more epigenetic modifications that are maintained in the embryo and endosperm after fertilization (Zhang et al., 2013a). The initiating mechanism leading to imprinting is poorly understood. These epigenetic modifications often reduce transcription levels of the imprinted allele that involves repressive histone marks (such as H3k27me3), cytosine DNA methylation and easiRNAs (Köhler and Weinhofer-Molisch, 2010; Zhang et al., 2013a; Borges et al., 2018; Martinez et al., 2018).

1.5.1 Controlling imprinted genes by DNA methylation

Studies analyzing the relationship between endosperm development and DNA methylation have shown that some methylated CG residues in the embryo are demethylated in the endosperm, leading to the demethylation in the endosperm being higher than in the embryo (Hsieh et al., 2009; Gehring et al., 2009). According to Hsieh et al. (2009), this is caused by the DNA glycosylase

DEMETER (DME), because the *dme* mutant was shown to partially restore DNA methylation. In addition, maternally imprinted genes in the vegetative tissue are expressed in the endosperm as a result of DNA demethylation (Köhler and Weinhofer-Molisch, 2010). Based on this hypothesis, a number of imprinted genes from the maternal genome have been identified in plants, including *MEDEA* (Choi et al. 2002), *FWA* (Kinoshita et al., 2004), *FIS2* (Jullien et al., 2006) and many more (Figure 6-A).

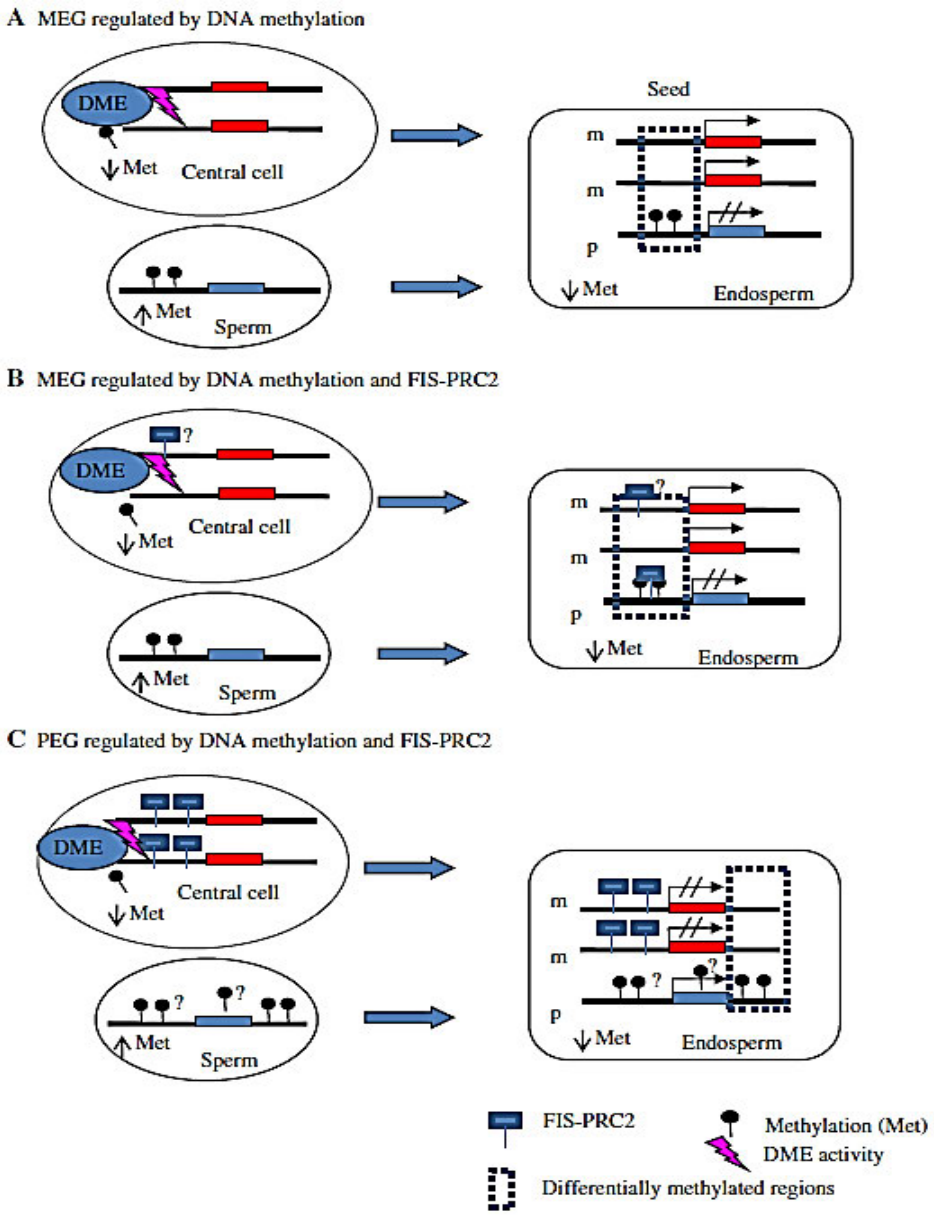


Figure 6. Model to explain the regulation of maternally and paternally expressed genes by DNA demethylation, methylation and FIS2–PRC2. In the central cell and endosperm, the activities of MET1, CMT3 and *de novo* DNA methyltransferases are low, but they are high in sperm. A: In the central cell, DME enzyme demethylates, leading to gene expression of the maternal allele (m).

The paternal allele (p) is methylated and silenced in sperm. For example, genes regulated by this mechanism include *FIS2* and *FWA*. B: DNA methylation and FIS2–PRC2 regulate MEGs. DME activity in the central cell leads to demethylation and gene expression of the maternal allele. In the central cell, the paternal allele is DNA methylated and modified with H3K27met catalysed by PRC2 in endosperm. However, DME may only be indirectly involved in imprinting, leaving PRC2 as the only mechanism. MEA is regulated by this mechanism. Whether the maternal allele is modified by H3K27met to modulate expression is unknown. C: DNA methylation and FIS2–PRC2 regulate PEGs. Using *PHE1* as an example, demethylation of the maternal allele by DME activity in the central cell leads to subsequent recruitment of PRC2 to the maternal allele, resulting in H3K27met and silencing. The paternal allele is DNA methylated and activated in sperm. DNA methylation occurs in downstream repeats at the *PHE1* locus (Zhang et al., 2013a).

The DNA methylation-based mechanism is not sufficient to explain the expression patterns of parent-specific genes (Figure 6-B and C). For example, the *PHERES1 (PHE1)* gene expressed in the paternal genome is imprinted in the endosperm by the repressive activity of PcG. However, repressing the *PHE1* maternal allele in endosperm relies on both the FIS-PcG complex binding to the promoter region of the *PHE1* locus and DME-mediated DNA demethylation at the 3' end of the *PHE1* locus (Makarevich et al., 2008; Hsieh et al*.,* 2009). Based on genome-wide analysis of imprinted genes in the endosperm, Hsieh et al. (2011) hypothesised that in the *fis* loss-of-function mutant, PEGs will be activated and expressed when fertilization occurs with *met1* pollen. However, the FIS-PcG targets needs to be demethylated so that the FIS-PcG complex can bind to it, meaning that methylation of alleles inherited from the paternal genome will prevent FIS targeting (Köhler and Kradolfer, 2011). Recent reports have shown that accumulation of easiRNAs at maternally imprinted loci likely mediated by RNA-directed DNA methylation (RdDM) activity (Martinez et al., 2018) bypasses hybridization barriers between diploid seed parents and tetraploid pollen parents in *A. thaliana* (Borges et al., 2018; Martinez et al*.*, 2018).

1.5.2 Polycomb group proteins control imprinted gene expression

Molecular-level studies of endosperm development have shown that the PRC2 complex plays an important role before and during endosperm development, as PRC2 mutants (like *fie, fis2* and *mea* mutants) display an autonomous endosperm phenotype before fertilization and later undergo additional

endosperm cell divisions and fail to undergo cellularization (Grini et al., 2002; Heo et al., 2013). The PcG complex is conserved through evolution and plays an important role in cell specification and organ development (Wang et al., 2004; Farrona et al., 2008). Two somewhat opposing models for targeting the PcG complex to chromatin have been proposed. One involves DNA-binding transcription factors binding to polycomb response elements (PREs) (He et al., 2013; Xiao et al. 2017) and the other involves noncoding RNAs (ncRNAs) acting as molecular guides (He et al., 2013; Borges et al., 2018; Martinez et al., 2018).

DNA-binding transcription factors directly or indirectly recruit the PcG complex by binding to the sequence-specific c*is* PREs and subsequently deposit repressive H3K27me3. One example, *AGAMOUS* (AG), a MADS-box transcription factor, has roles in repressing the *WUSCHEL (WUS)* locus by binding to a CArG sequence at the *WUS* locus and thus directly or indirectly recruiting the PRC2 complex and LHP1 (LIKE HETEROCHROMATIN PROTEIN1) (Liu et al., 2011). Other examples of DNA-binding transcription factors recruiting the PcG have been described in *A. thaliana* by Xiao et al. (2017). These authors identified two transcription factors, AZF1 (AZOOSPERMIA FACTOR 1) and BPC1 (BASIC PENTEACYSTEINE 1), that bind to the short genomic fragments known as PREs that contain a GA-repeat motif and a telobox motif to co-localize with PRC2 on chromatin and physically interact with and recruit PRC2 to the target genes. All of these suggest that DNA-binding transcription factors play important roles in targeting the PcG complex.

In animals, ncRNAs have been shown using the *cis* or *trans*-acting method to have important roles in recruiting PRC2 complex to target sites (Beisel and Paro, 2011). In plants, the function of ncRNAs in PRC2 recruitment is unclear. Recently, two studies have suggested that easiRNAs might play a role in recruiting the PRC2 complex to the targets in *A. thaliana*. They showed that a highly conserved microRNA in plants, miR845, targets the $tRNA^{Met}$ primer-binding site of long terminal repeat retrotransposons, triggering the accumulation of 21-22-nt small RNAs in a dose-dependent fashion via RNA polymerase IV, leading to PRC2 recruitment (Borges et al., 2018) or RdDM activity (Martinez et al., 2018) at maternally imprinted loci, which helps bypass the triploid block in response to

increased paternal ploidy in *A. thaliana*. However, the detailed mechanism for involvement of easiRNAs in silencing MEGs by PRC2 recruitment is still not clear.

lncRNAs can also directly or indirectly recruit the PcG complex. The mammalian PcG complex binds to hundreds of lncRNAs that are thought to act as sequence-specific guides directing the complex to the chromatin to deposit repressive histone (H3K27me3) marks (Khalil et al., 2009; Beisel and Paro, 2011; Davidovich and Cech, 2015). The best-described example in plants is at the *FLOWERING LOCUS C* (*FLC*) locus in *Arabidopsis*. Two lncRNAs regulating *FLC* expression have been identified: *COLD-INDUCED LONG ANTISENSE INTRAGENIC RNA* (*COOLAIR*), and *COLD-ASSISTED INTRONIC NONCODING RNA* (*COLDAIR)* (Figure 7).

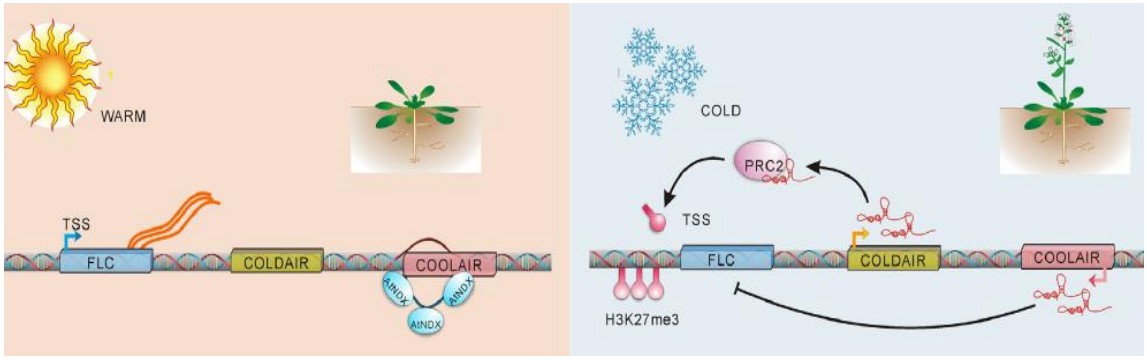


Figure 7. PRC2 recruitment by lncRNAs. Schematic representation of the roles of COLDAIR and COOLAIR on *FLC* expression and the regulation of COOLAIR by AtNDX during the course of vernalization (Zhang et al., 2013a).

COOLAIR is an antisense RNA that is transcribed in response to cold treatment and is alternatively polyadenylated at the 3' end, resulting in a proximal poly(A) site or a distal poly(A) site (Swiezewski et al., 2009, Liu et al., 2010). The proximal poly(A) site stimulates the activity of *FLD*, a homologue of the human LYSINE SPECIFIC DEMETHYLASE1 (LSD1; Sanda and Amasino, 1996; Liu et al., 2007), to reduce the level of H3K4me2 at the *FLC* locus, leading to a transition from an active chromatin state to a repressive state (Liu et al., 2010). Reduction in H3K4me2 might benefit H3K27me3 modification; thus, *COOLAIR* acts as an indirect recruiter of PRC2. However, how *FLD* is activated using the proximal site of *COOLAIR* remains unknown.

COLDAIR is a sense ncRNA that has a 5' cap but no poly (A) tail, and is induced at low temperatures (Heo and Sung, 2011). *COLDAIR* can directly interact with the CXC domain of the core PcG components. In *COLDAIR* knockdown plants, the PcG complex is not properly recruited to *FLC*, resulting in insufficient H3K27me3 modification at the *FLC* locus. Therefore, COLDAIR serves as a direct recruiter for PcG.

Collectively, these lines of evidence were sufficiently strong to support an observation: PRC2 binds RNA. However, a central role for lncRNAs as a major driving force in the recruitment of PRC2 in a gene-specific manner is still not clear, as exciting as it would be for those of us immersed in the RNA World.

1.6 Long Noncoding RNAs

lncRNAs are mainly transcribed by DNA-dependent RNA polymerase II (Pol II), are sometimes polyadenylated, often spliced and mostly localized within the nucleus (Wierzbicki, 2012; St Laurent et al., 2015). Along with RNA Pol II-derived lncRNAs, in plants, Pol IV also transcribes thousands of lncRNAs, but these are co-transcriptionally processed into double-stranded RNA (dsRNA) and then into small interfering RNA (siRNA) (Shin and Shin, 2016). While the full repertoire of lncRNA functions in plants is still to be elucidated, they have a key role in flowering time regulation (Leeuwen and Mikkers, 2010; Heo and Sung, 2011) and responses to pathogen invasion (Xin et al., 2011; Liu et al., 2012; Seo et al., 2017) and are transcribed in a changing environment (Kruszka et al., 2012; Wang et al., 2017). Generally, there are four main mechanistic themes or archetypes of lncRNA activity, as shown in Figure 8 (Rinn and Chang, 2012).

1.6.1 Classification of long noncoding RNAs

In mammalian and plant genomes, most transcribed genomic regions encode noncoding RNAs. These ncRNAs are typically transcribed from introns, antisense to exons, intergenic and often do not yet have any prescribed biological function (Figure 8). ncRNAs fall into two broad groups based on their size: small ncRNAs shorter than 200 nt, such as microRNAs (miRNAs), siRNAs, *trans*-acting siRNAs (tasi-RNAs), small nucleolar RNAs (snoRNAs) and transfer RNAs (tRNAs); and

lncRNAs, which are longer than 200 nt and up to 100 kp in animals (Czech and Hannon 2011; Siomi et al., 2011; Bai et al., 2014; St Laurent et al., 2015). Small and long ncRNAs play important roles in regulating biological processes such as cell differentiation during development and metabolism (Mercer et al., 2009).

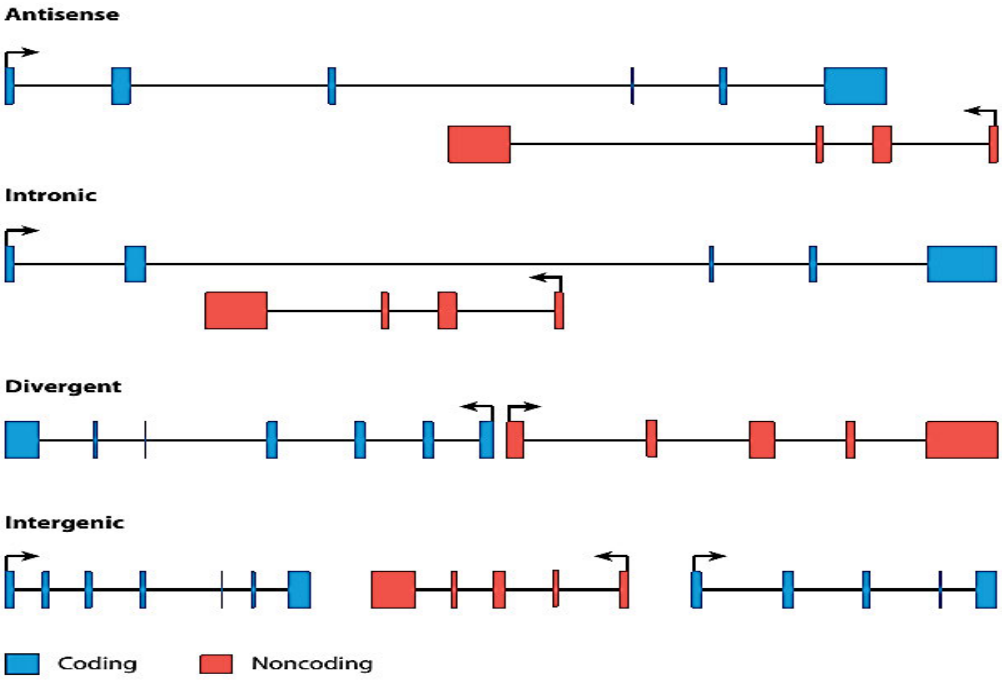


Figure 8. Genomic locations of lncRNAs in relation to protein-coding genes. Antisense lncRNA transcription initiates inside or at the 3' end of protein-coding genes; lncRNA transcription is opposite to a protein-coding gene and overlaps with at least one of the exons of the protein-coding gene. Intronic lncRNAs are located inside an intron of a protein-coding gene in either the sense or antisense direction, and their ends do not have any overlap with exons. Divergent lncRNAs are transcribed in the opposite direction to that of a nearby protein-coding gene. Intergenic lncRNAs (also called large intervening ncRNAs or lincRNAs) are transcribed from loci localized between protein-coding genes (Rinn and Chang, 2012).

1.6.2 Molecular roles of long noncoding RNAs

The exact mechanism of lncRNA function is still unclear. To date, several mechanisms have been hypothesised: (1) RNA–DNA–DNA triplex (*trans*-); (2) RNA–DNA hybrid; (3) RNA–RNA hybrid of lncRNA with a nascent transcript; and (4) RNA–protein interaction (*cis*-/*trans*-). However, only (1), (2) and (4) have been demonstrated and experiments have shown that lncRNAs interact with partners such as DNA, RNA or protein to carry out their functions as decoys, scaffolds,

guides and enhancers (Figure 9). Examples are given below of these mechanisms.

First archetype, lncRNAs can act as decoys that indirectly inhibit regulatory proteins by preventing their association with their target DNA. Two examples of this in animals are lncRNA GROWTH ARREST-SPECIFIC 5 (GAS5), which binds to the glucocorticoid receptor at the DNA-binding site to prevent it interacting with DNA, stopping metabolic gene transcription (Kino et al., 2010); and the blocking of components in the silencing machinery to prevent progression of the cycle, such as lncRNA P21-ASSOCIATED ncRNA DNA DAMAGE-ACTIVATED (PANDA), which binds to transcription factor NF-YA, preventing the apoptosis mediated by p53 (Hung et al., 2011).

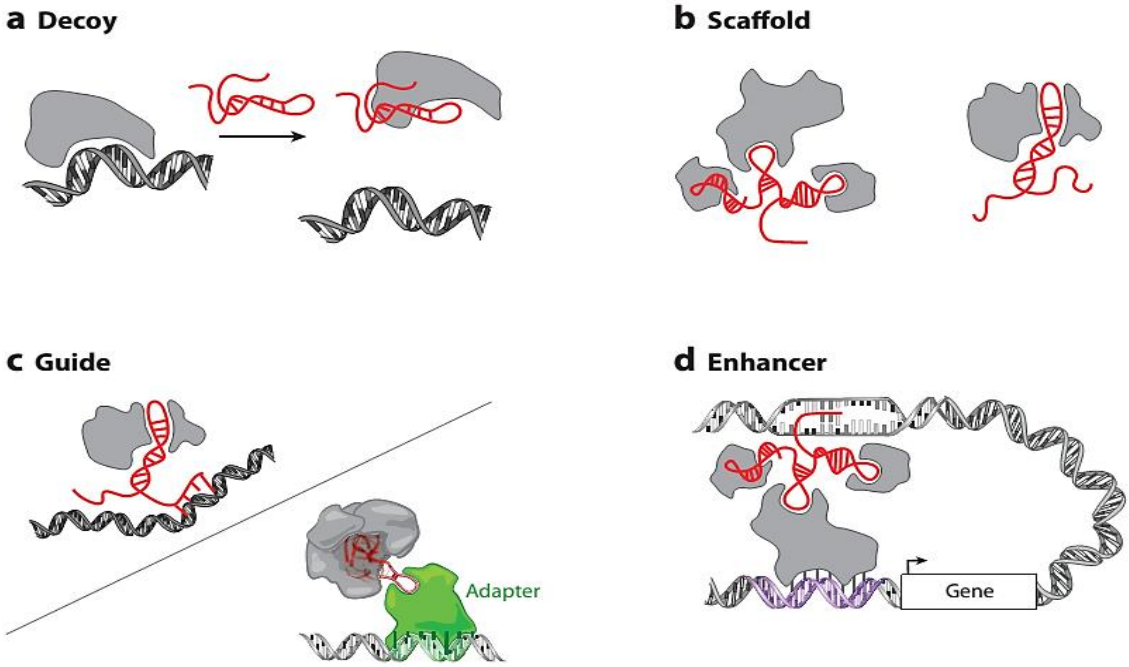


Figure 9. Mechanistic themes of lncRNA activity: (a) as decoys, lncRNAs can bind microRNAs or target proteins such as transcription factors and titrate away the DNA-binding protein; (b) as a scaffold for recruitment of proteins into a ribonucleprotein complex; (c) as a guide to recruit chromatin or DNA-modifying enzymes to their target histone or DNA region; (d) as enhancers to primarily enhance DNA-dependent RNA polymerase II transcription (Rinn and Chang, 2012).

The second archetype of lncRNA function is as a central platform for the assembly of molecular components (like proteins, peptides) into a complex (Spitale et al., 2011). For example, the lncRNA HOTAIR has specific sites in its

structure that help it simultaneously bind to two complexes, PRC2 and LSD1-CoREST, to control gene silencing (Tsai et al., 2010).

The third archetype of lncRNA function is as molecular guides, whereby they bind to protein complexes to facilitate their localization to specific targets (Rinn and Chang, 2012). For example, in *Arabidopsis*, lncRNA COLDAIR binds to the PRC2 complex to guide deposition of H3K27me3 to chromatin at *FLC* to enable the induction of flowering (Heo and Sung, 2011) (see Figure 6 for summary).

The fourth archetype function of lncRNAs is as enhancers, particularly for transcriptional enhancement. For instance, the nascent RNA of the mammalian lncRNA *HOTTIP* creates a chromosomal loop when *HOTTIP* appears at specific *Hox* loci, which activates the transcription of its target gene (Wang et al., 2011).

In summary, although lncRNAs can be classified into four types as above, more archetypes most likely exist as the majority of transcribed lncRNAs do not have an associated biological function that may be detected in future.

1.7 Long Noncoding RNAs Modulate the Activity of RNA-binding Protein Complex and Regulate Gene Expression

Regulation of gene expression plays an important role in many complicated processes in the body, such as development, differentiation, cell specification and responses to environmental stimuli. Post-transcriptional regulatory processes have critical effects on eukaryotic gene expression programs like pre-mRNA splicing and maturation, as well as mRNA transport, stability, storage, editing, translation and turnover. ncRNAs including miRNAs, lncRNAs, together with RNA-binding proteins (RBPs), play an important role in such processes (Fabian et al., 2010; Morris et al., 2010; Castello et al., 2012; Yoon et al., 2013; Moore, 2015).

RBPs have been shown to have important roles in regulating many cellular processes in plants and animals, such as cell proliferation, death, differentiation and development (Yang et al., 2015; Wang and Chekanova, 2017) (Table 2). In addition, RBPs can lead to differential expression or altered activity of certain RBPs that are involved in the pathogenesis of several human diseases (Keene,

2007; Lukong et al., 2008). RBPs primarily bind to specific sequence elements in newly synthesized or mature RNAs to regulate their expression by affecting pre-mRNA splicing and maturation as well as mRNA transport, storage, turnover and translation (Lukong et al., 2008; Zhang et al., 2013a). Depending on the type of RBP and the associated RNA sequence, RBPs can bind to RNA through an RNA-recognition motif (RRM) or RNA-binding domain (RBD) in either the nucleus or the cytoplasm. For example, Lorkovic and Barta (2002) reported 196 *Arabidopsis* RBPs with the RRM and 26 RBPs with K-homology; Miller et al. (2008) reported that the Pumilio/FBF (PUF) family proteins have other RBDs like the PUF repeat. These domains interact with associated RNA sequences in a specific or non-specific manner. Bailey-Serres et al. (2009) identified over 1,100 RBPs in *A. thaliana*, of which 200 are functionally characterized as RBPs involved in canonical processes of splicing and translation. Hence, studying RNA–RBP networks and RNA sites bound by RBPs is important for fully understanding the complex regulatory processes in the body, and will likely provide supporting evidence that RBPs are important for cellular regulation.

Recently, the significance of lncRNA–protein interactions has been better understood with respect to molecular mechanisms in some biological processes (Table 2). However, the biochemical attributes of these interactions are being discovered and novel bioinformatics approaches are being developed to identify and predict proteins that interact with target lncRNA, and vice versa. RNA immunoprecipitation (RIP) is a technique that depends on the fixation of samples to cross-link RBPs to RNAs *in vivo*, followed by immunoprecipitation of specific ribonuclear protein (RNP) complexes and identification of associated RNAs. The advantages of RIP are that the cross-linking and denaturing conditions during extract preparation and the immunoprecipitation step minimise a recognised problem in standard immunoprecipitation: the re-association of RBPs with non-cognate RNAs occurring in cellular extracts (Mili and Steitz, 2004). Importantly, these conditions do not affect native RNA–protein complexes because they are stabilized by the cross-linking of their components. However, this method also has some limitations, including that (1) the antibody used and the abundance of the target ribonucleoprotein strongly affect the results; and (2) RNA molecules

are known to be ‘sticky’ and might exhibit non-specific binding to RBPs (Niranjanakumari et al., 2002).

The RIP assay has been used to analyze native RNA–protein interactions of plant RBP complexes (Terzi and Simpson, 2009; Köster and Staiger, 2014; Sorenson and Bailey-Serres, 2015). These methods use a cell-lysis buffer that is likely able to stabilize different RNPs for immunoprecipitation.

Table 2. Summary of lncRNA binding proteins in plants and animals

	LncRNA	RBP	Biological function	Reference
Animals	ANRIL	PRC2, PRC1	Affecting *p16*^{INK4a} gene expression and cell senescence	Wang et al., 2004
	AIR	G9a	Targeting G9a in *cis* for imprinting	Nagano et al., 2008
	FENDRR	PRC2, WDR5	Regulating genes in *cis* and in *trans*	Grote et al., 2013
	FIRRE	hnPNPU	Modulating the nuclear architecture across chromosomes	Hacisuleyman et al., 2014
	HOTAIR	PRC2, LSD1	Silencing transcription in *trans* via its modular architecture	Tsai et al., 2010
	HOTTIP	MLL-WDR5	Activating gene expression via chromosomal looping	Wang et al., 2011
	lincRNA-p21	hnRNP-K	Mediating p53-dependent gene repression	Huarte et al., 2010
	MALAT1	PSPC1, PSF, PURA	Sequestering splicing factor to regulate alternative splicing	West et al., 2014
	NEAT1	PSPC1, SRSF1, ESRP2	Playing a role in RNA processing and transcriptional regulation	West et al., 2014
	Rox1	MLE, MSL	Mediating X chromosome upregulation to rescue male lethality	Quinn and Chang, 2015
	TERC	TCAB1	Having functions as the template and scaffold for the telomerase complex	Chu et al., 2011
	Xist	81 proteins (Hnrnpk, Spen)	Mediating chromatin modification and polycomb targeting	Chu et al., 2015
	Xist	10 proteins (SHARP, HDAC3)	Interacting directly with SHARP to silence transcription through HDAC3	McHugh et al., 2015
Plants	ASCO-lncRNA	NSR	Having functions in lateral root development in *Arabidopsis*, as a regulator of alternative splicing and as a decoy lncRNA	Bardou et al., 2014

	LncRNA	RBP	Biological function	Reference
	COLDAIR	PRC2	Having functions in regulation of flowering in *Arabidopsis* in the vernalization pathway, showing an association with polycomb to mediate silencing of *FLC* and affecting chromatin looping at *FLC* in response to vernalization	Heo and Sung, 2011
	COLDWRAP	PRC2	Having functions in regulation of flowering in *Arabidopsis* in the vernalization pathway; participating in and coordinating vernalization-mediated polycomb silencing of the *FLC*; also having an effect on formation of an intragenic chromatin loop that represses *FLC*	Kim and Sung, 2017
	ANTISENSE LONG (ASL)	RRP6L1	Regulating flowering in the autonomous pathway in *Arabidopsis*; AtRRP6L controls ASL to modulate H3K27me3 levels	Shin and Chekanova, 2014
	APOLO	AGO4	Having functions in regulation of auxin signalling outputs in *Arabidopsis*; participating in chromatin loop dynamics; influencing formation of a chromatin loop in the PINOID promoter region	Ariel et al., 2014
	Pol V transcripts	AGO4	Having roles in silencing TEs and repeats in RdDM pathway; also, serving as a scaffold lncRNA for assembly of siRNAs and proteins in the RdDM pathway	Böhmdorfer et al., 2016
	ENOD40	NSR	Regulating symbiotic interactions between leguminous plants and soil bacteria in *Medicago truncatula*; possible function in re-localization of proteins in plants	Bardou et al., 2014
	ELENA1	MED19a	Having roles in protecting the plant against the P*seudomonas syringae* pv. *tomato* DC3000	Seo et al., 2017

1.8 Long Noncoding RNA Associates with Transposable Elements to Regulate Gene Expression

With the development of high-throughput sequencing technology, identification of ncRNAs has been extensively described in plant and animal transcriptomes. Among these ncRNAs, a growing number of lncRNAs has been identified in multicellular organisms; however, the precise functions as well as the origin and evolution of many lncRNAs remain to be explored (Chitwood and Timmermans, 2010; Liu et al., 2012; Chen et al., 2015). In many eukaryotic genomes, transposable elements (TEs) are mobile genetic elements with many copies, widely distributed and often accounting for a large fraction of plant and animal genomes (de Koning et al., 2011; Liu et al., 2012). TEs are increasingly recognised as important players in the origins of functional novelties (Feschotte, 2008). Several instances of TEs have revealed that they can be a source of *cis* elements regulating expression of adjacent genes (Kunarso et al., 2010). TEs have also been reported as major factors in the expression of miRNA genes (Li et al., 2011; Zhou et al., 2013). Further, TEs have been found to be remarkably enriched within lncRNA exons relative to protein-coding exons (Kelley and Rinn, 2012; Kannan et al., 2015). Notably, Chishima et al. (2018) found that many TE–tissue pairs are associated with tissue-specific expression of lncRNA in humans, and suggested that multiple TE families can be re-used as functional domains or regulatory sequences of lncRNAs. In addition to functions of TE-associated lncRNAs, a recent study showed that TE-associated lncRNAs play an important role in plant biotic stress responses in *A. thaliana* (Wang et al., 2017). All of these findings support the hypothesis that TEs might serve as one of the functional elements in lncRNAs.

1.9 Context of This Study

To date, the role of lncRNAs in seed development remains unclear. In this research project, lncRNAs were identified in *A. thaliana*, rice and maize in an attempt to discover lncRNAs involved in early seed development.

1.10 Aims of This Project

The specific aims of this project were as follows.

1.10.1 Establish methodology for identification and purification of plant long noncoding RNAs (Chapter 2)

This study optimised experimental conditions to purify specific cell types, undertook bioinformatic annotation of lncRNAs and investigated biological function using the reference plant *A. thaliana*.

1.10.2 Establish methodology for quantitative and single-nucleotide resolution profiling of RNA 5-methylcytosine (Chapter 3)

This study developed methods for highly reproducible bisulfite treatment of RNA, efficient locus-specific PCR amplification, detection of candidate sites by sequencing on the Illumina MiSeq platform and bioinformatic calling of non-converted sites.

1.10.3 Explore the contribution of transposable elements to intergenic long-noncoding RNAs (Chapter 4)

In many eukaryotic genomes, TEs are widely distributed; they often account for large fractions of plant and animal genomes. However, the contribution of TEs to lincRNAs is largely unknown (Gregory, 2005; de Koning et al., 2011). By using strand-specific RNA sequencing, the expression patterns of TE-associated lincRNAs in *Arabidopsis*, rice and maize were profiled.

1.10.4 Identify long noncoding RNAs bound to the FIS2–PRC2 complex (Chapter 5)

In mammals, around 9,000 lncRNAs bind to PRC2 (Khalil et al., 2009). However, only two plant lncRNAs have been demonstrated to bind to PRC2 (Heo and Sung, 2011; Kim and Sung, 2017). Hence, identification of lncRNAs bound to the FIS2–PRC2 complex is predicted to reveal novel lncRNAs involved in seed development. To identify FIS2–PRC2-associated lncRNAs, transcriptome-wide

RNA sequencing of *A. thaliana* siliques was performed on the next-generation sequencing platform, HiSeq 2000 (Illumina®).

1.10.5 Identify long noncoding RNAs in *Arabidopsis thaliana* endosperm (Chapter 6)

To identify lncRNAs, transcriptome-wide RNA sequencing of *A. thaliana* endosperm was performed on the next-generation sequencing platform, HiSeq 2000 (Illumina®).

Chapter 2: Purification and Functional Analysis of Plant Long noncoding RNAs (lncRNAs)

Trung Do[1], Zhipeng Qu[1] and Iain Searle[1,*]

[1]School of Biological Sciences, Department of Molecular and Biomedical Sciences, The University of Adelaide, Australia

*Corresponding Author, iain.searle@adelaide.edu.au

Submitted for a chapter in the Methods in Mol. Biol. book on Plant lncRNAs.

Statement of Authorship

Title of paper	Purification and functional analysis of plant long non-coding RNAs (lncRNAs)
Publication Status	Accepted
Publication details	In press Spinger Science + Business, Media, LLC, New York

Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. The candidate's stated contribution to the publication is accurate (as detailed below)
ii. Permission is granted for the candidate in include the publication in the thesis and
iii. The sum of all co-author contributions is equal to 100% less the candidate's state contribution.

Name of Candidate	Trung Q. Do		
Contribution to chapter	Wrote the manuscript and composed figures.		
Overall percentage (%)	80%		
Signature		Date	15/03/2018

Name of Co-Author	Zhipeng Qu		
Contribution to chapter	Wrote the Materials and Methods section on bioinformatics analysis. Overall contribution 5%.		
Signature		Date	15/03/2018

Name of Co-Author	Iain R. Searle		
Contribution to chapter	Supervised development of work and edited the manuscript. Overall contribution 15%.		
Signature		Date	15/3/18

Abstract

More than 70% of eukaryotic genomes are transcribed into RNA transcripts, the majority of these transcripts are noncoding protein and their biological functions are largely unknown. Over the last decade, the application of high throughput sequencing technologies has led to the description of almost all cellular coding and noncoding RNA transcripts except perhaps for those transcripts that are lowly abundant or those present only in specific cells that are underrepresented in sampled tissue(s). An often under represented class of noncoding are long noncoding RNAs (lncRNAs) and these often play key regulatory functions for many biological processes such as cell identity and cell division. However, the purification and functional characterization *in vitro* is still a challenge in both animal and plant experimental systems. Here, we describe in detail methodology for purification of specific cell types, bioinformatic annotation of lncRNAs and investigation of biological function using the reference plant *Arabidopsis thaliana*.

Keywords

Arabidopsis thaliana, functional analysis, long noncoding RNA, nuclei purification, RNA-Seq.

1 Introduction

Eukaryotic genomes transcribe genetic information from chromatin, into RNA and subsequently a small portion is translated into proteins. However, the majority of these RNA transcripts are not translated, and are described as noncoding RNAs (ncRNAs) ***(1, 2)***. Many ncRNAs are post-transcriptionally processed, examples include introns removed, covalent RNA modifications added or diced into smaller

ncRNAs (***3-6)***. Arbitrarily, ncRNAs have been grouped by size into small, sRNAs, those less than 200 nucleotides (nt) and long ncRNAs (lncRNAs) those larger than 200 nt in length ***(7-9)***. To date, lncRNAs have been shown to have diverse functional roles in many fundamental cellular processes and they also represent important components of ongoing research in many fields ***(10)***.

Often the first step is to identify or sequence the lncRNAs within specific tissue(s) of interest. However, cell-specific expression or low abundance of the lncRNA within a tissue can often make this first step a challenge. In this chapter, we adapt and describe a method called isolation of nuclei tagged in specific cell types (INTACT) to purify a specific cell type (***11***). The next step is often application of sequencing technology to sequence the purified RNA and bioinformatic annotation of protein coding and noncoding transcripts. This method has already been applied to a large number of species including *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, mouse and human cells ***(12-21)***. We briefly describe these bioinformatic steps within the chapter. Often the next step is functional characterization of novel lncRNAs and this can involve perturbing transcription by CRISPR mediated deletion, overexpression of the lncRNA by using a transgene or transcriptional activators, or knockdown using dsRNA or artificial miRNAs ***(22-25)***. In this chapter, we detail protocols for the functional characterization of lncRNAs from *Arabidopsis thaliana* siliques (**Fig. 1**).

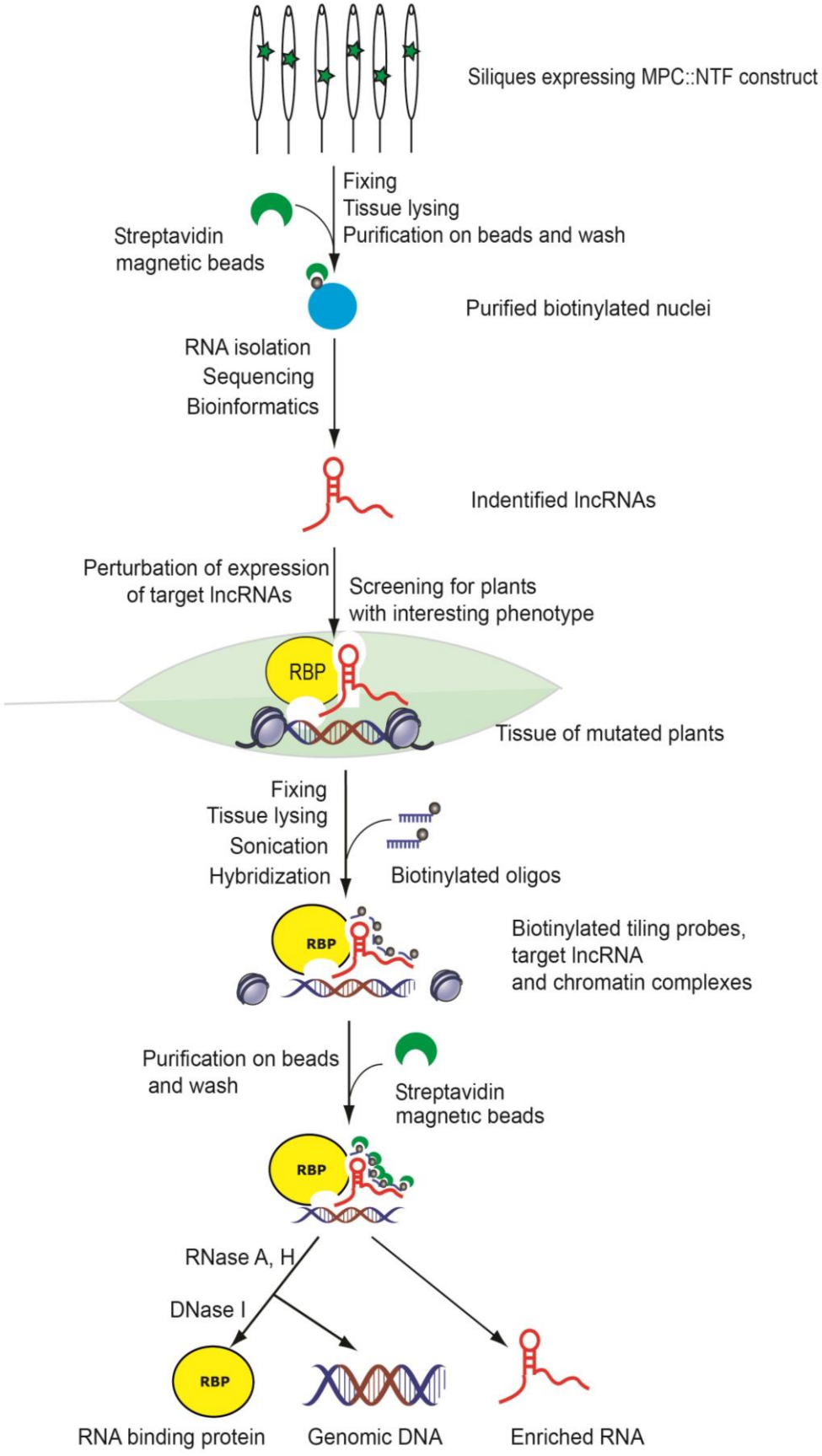

Siliques expressing MPC::NTF construct
Fixing
Tissue lysing
Purification on beads and wash
Streptavidin
magnetic beads
Purified biotinylated nuclei
RNA isolation
Sequencing
Bioinformatics
Indentified lncRNAs
Perturbation of expression
of target lncRNAs
Screening for plants
with interesting phenotype
RBP
Tissue of mutated plants
Fixing
Tissue lysing
Sonication
Hybridization
Biotinylated oligos
RBP
Biotinylated tiling probes,
target lncRNA
and chromatin complexes
Purification on beads
and wash
Streptavidin
magnetic beads
RBP
RNase A, H
DNase I
RBP
RNA binding protein
Genomic DNA
Enriched RNA

Fig. 1 Overview showing the workflow for this chapter. The siliques of transgenic plants expressing the MPC::NTF construct (green stars) were collected and fixed with the formaldehyde. The fixed samples were then lysed into purification buffer and streptavidin magnetic beads were used to enrich the biotinylated nuclei. Next, RNA was extracted from purified nuclei, libraries constructed and sequenced using an Illumina Nextseq. The data were analyzed by a bioinformatics pipeline to identify potential long noncoding RNAs (lncRNAs). The target lncRNAs were confirmed and functional analysis undertaken by perturbing their expression by using an overexpression or knockdown vector. Mutants with an interesting phenotype were used as material for chromatin isolation by RNA purification (CHIRP) experiments to identify the target genomic DNA and/or target-lncRNA binding proteins.

2 Materials

Prepare all solutions using RNase-free and DNase-free H_2O and analytical grade reagents. Store and prepare all reagents at room temperature unless indicated otherwise. Prepare and perform RNA extraction, cDNA synthesis, and PCR amplification experiments in an RNase-free area. Follow all state or national safety and waste disposal regulations when performing experiments.

2.1 Plant growth and tissue sampling

1. Transgenic seeds expressing the *E. coli* biotin ligase BirA and tissue specific nuclear targeting fusion protein (NTF) in the endosperm
2. Appropriate plant growth media
3. Liquid nitrogen (and container)
4. Polystyrene box and/or second liquid nitrogen-proof container
5. Sharp knife, scalpel, razor blade, tweezers, metal needle/probe and flame source
6. Eppendorf tubes (1.5 mL)

7. –80°C freezer or liquid nitrogen storage container and/or dry ice
8. RNA*later*™ solution (ThermoFisher Scientific)

2.2 Nuclei purification

2.2.1 Materials and reagents

1. 37% (w/v) Formaldehyde
2. Glycine
3. Liquid nitrogen (N_2)
4. MOPS (Merck)
5. NaCl
6. KCl
7. EDTA
8. EGTA
9. Spermidine
10. Spermine
11. cOmplete™, Mini, EDTA-free Protease Inhibitor Cocktail (Merck)
12. Tris_HCl
13. SDS
14. 4′,6-diamidino-2-phenylindole (DAPI)
15. M-280 streptavidin Dynabeads (ThermoFisher Scientific)
16. Triton X-100
17. 40-µM cell strainer (BD Falcon)
18. MiniMACS™ separation magnet (Miltenyi Biotec)
19. Refrigerated microcentrifuge (Eppendorf, model 5415R or equivalent)
20. Refrigerated centrifuge (Sorvall, model RC5C or equivalent)

21. Rotating mixer for 1.5-mL tubes (Labquake; ThermoFisher Scientific, or equivalent)
22. Electronic serological pipetting device (Easypet; Eppendorf, or equivalent)
23. Fluorescence microscope with DAPI and GFP filters

2.2.2 Nuclei purification buffers (NPB)

Spermidine, spermine and cOmplete™ protease inhibitors are added just before use and the solution is kept on ice.

1. 20 mM MOPS (pH 7.0)
2. 40 mM NaCl
3. 90 mM KCl
4. 2 mM EDTA
5. 0.5 mM EGTA
6. 0.5 mM spermidine
7. 0.2 mM spermine
8. 1× cOmplete™ protease inhibitors

2.2.3 Nuclear lysis buffer

This solution should be prepared just before use and kept at RT. Do not store this solution.

1. 50 mM Tris (pH 8.0)
2. 10 mM EDTA
3. 1% (w/v) SDS
4. 1× cOmplete™ protease inhibitors

2.3 RNA extraction

1. TRIzol™ reagent (ThermoFisher Scientific). Refer to the manufacturer's instructions and guidelines for stability and storage, and handle with eye and glove protection.
2. Chloroform (ThermoFisher Scientific)
3. Isopropyl alcohol
4. Absolute ethanol
5. RNase-free water
6. 0.5% SDS and UV-treated plastic ware, oven-baked sterile glassware
7. Eppendorf tubes, or clean sterile Falcon tubes (conical bottom)
8. Liquid nitrogen, mortar and pestle
9. Benchtop centrifuges (refrigerated or access to cold room)

2.4 Generation of RNA-Seq library

1. Ribo-Zero magnetic kit (Illumina)
2. NEBNext® Ultra™ Directional RNA Library Prep Kit (Illumina)

2.5 Library sequencing

1. 0.2 M NaOH
2. 0.2 M Tris-HCl, pH 7.0
3. NextSeq 500 Kit
4. Illumina PhiX Control Kit

2.6 Bioinformatics analysis

1. FASTQ RNA-Seq files

2. Pipeline to identify the lncRNAs

2.7 Functional analysis of lncRNA

1. Transcripts of target lncRNAs
2. *Arabidopsis thaliana* seeds
3. Donor and binary vectors (pENTR/D, pLeela, pJawohl18, pRS300, pGreen II)
4. Appropriate antibiotics
5. Plant growth facilities
6. Primers of target lncRNAs (*LNCRNA_1246*_R: 5'-TGACCTGCTGCTCTCATCTCG-3', *LNCRNA_1246*_F: 5'-GTTGCACATCAGGGACATG-3'), and house-keeping genes (Actin1_F: 5'-GTCTCGAGAGATGACTCAGATCATGTTTGAG-3'; Actin1_R: 5'-GGCGCGCCACAATTTCCCGTTCTGCGGTAG-3')

2.8 Chromatin isolation by RNA purification (ChIRP)

2.8.1 Materials and reagents

1. T_3 seed of target-lncRNA mutants
2. Agar
3. Phosphate buffered saline (PBS)
4. Formaldehyde
5. 0.125 M glycine
6. 40 µM strainer (BP Falcon)
7. Vacuum indicator (Sorvall, model RC5C or equivalent)

8. Probes, design an array of probes along the lncRNAs by using Stellaris® Probe Designer version 4.2 at http://www.singlemoleculefish.com
9. Yeast total RNA (Merck)
10. Bovine serum albumin (BSA)
11. DNase I (ThermoFisher Scientific)
12. RNA extraction kit: RNeasy Mini column extraction (other commercial RNA extraction kits can be used) (Merck)
13. $MgCl_2$
14. Sucrose
15. Beta-mercaptoethanol
16. Dithiothreitol (DTT)
17. Phenylmethylsulfonyl fluoride (PMSF)
18. RNaseOut™ (ThermoFisher Scientific)
19. $NaHCO_3$
20. NP-40 (Merck)

2.8.2 Buffers

1. Buffer 1: 0.4 M sucrose, 10 mM Tris-HCl, pH 8.0, 10 mM $MgCl_2$. Before use, add 5 mM beta-mercaptoethanol, 1 mM DTT, 1 mM PMSF, 1 tablet of cOmplete™ protease inhibitor and 0.1 U/μL RNaseOut™.
2. Buffer 2: 0.25 M sucrose, 10 mM Tris-HCl, pH 8.0, 10 mM $MgCl_2$, 1% Triton X-100. Before use, add 5 mM beta-mercaptoethanol, 1 mM DTT, 1 mM PMSF, 1 tablet of cOmplete™ protease inhibitor and 0.1 U/μL RNaseOut™.

3. Buffer 3: 1.7 M sucrose, 10 mM Tris-HCl, pH 8.0, 2mM $MgCl_2$, 0.15% Triton X-100. Before use, add 5 mM Beta-mercaptoethanol, 1 mM DTT, 1 mM PMSF, 1 tablet of cOmplete™ protease inhibitor and 0.1 U/μL RNaseOut™.
4. Lysis buffer: 50 mM Tris-HCl, pH 8.0, 10 mM EDTA, 1% SDS. Before use, add 5 mM beta-mercaptoethanol, 1 mM DTT, 1 mM PMSF, 1 tablet of cOmplete™ protease inhibitor and 0.1 U/μL RNaseOut™.
5. Hybridization buffer: 500 mM NaCl, 1% SDS, 100 mM Tris, pH 7.0, 10 mM EDTA, 15% formamide. Before use, add 5 mM beta-mercaptoethanol, 1 mM DTT, 1 mM PMSF, 1 tablet of cOmplete protease inhibitor and 0.1 U/μL RNaseOut™.
6. Wash buffer: 2xSSC, 0.5% SDS. Add 1 mM DTT and 1 mM PMSF fresh
7. RNA elution buffer: Tris-HCl, pH 7.0 and 1% SDS
8. DNA elution buffer: 50 mM $NaHCO_3$, 1% SDS, 200 mM NaCl
9. DNase buffer: 100 mM NaCl and 0.1% NP-40

3 Methods

Carry out all procedures described below at room temperature unless otherwise stated.

3.1 Sampling

Collect plant material (1 day after pollination siliques, DAP) from plant grown in a controlled environment facility, place the siliques in pre-weighed Eppendorf tubes and immediately immerse in liquid nitrogen. Alternatively, submerge the siliques in RNA*later*™ solution (*see* **Notes 1-5**).

3.2 Purification of nuclei from specific plant cell types using the INTACT method

1. In a mortar and pestle pre-cooled with liquid N_2, grind 5g of siliques and re-suspend the tissue powder in 10 mL of cold NPB buffer. Keep on ice for 10 minutes (*see* **Note 6**).
2. Filter the extract through a 40 µM strainer and centrifuge down the nuclei at 1200g for 15 minutes at 4°C.
3. Discard the supernatant and gently re-suspend the nuclei in 1 mL of cold NPB and transfer to a 1.5 mL tube.
4. Wash the appropriate amount (25 uL of beads for each 5g of root tissue or 10 uL for each 0.5g of leaf tissue) of Dynabead™ M-280 Streptavidin beads with 1 mL of NPB and then re-suspend the beads with NPB to their original volume. Add the bead suspension to the nuclei from Step 3 and rotate at 4°C for 30 minutes.
5. Dilute 1 mL of bead-nuclei mixture with 14 mL of 4 ice cold NPB containing 0.1% Triton X-100 (NPBt) in a 15 mL tube. Mix gently and place on ice for 30 seconds. Place the tube in the DynaMag™-5 magnet for 5 minutes at 4°C.
6. Carefully remove the supernatant with a serological pipette and gently re-suspend the beads in 14 mL of cold NPBt. Mix gently and place on ice for 30 seconds. Place the tube in the DynaMag™-5 magnet for 5 minutes at 4°C.
7. Repeat step 5.

8. Gently remove the supernatant with a serological pipette and resuspend the beads in 1 mL of cold NPBt. Remove a 25 uL sample for counting the number of captured nuclei on a hemocytometer.
9. Transfer the resuspended beads to a 1.5 mL tube and capture on a DynaMag™-5 magnet.
10. Remove the supernatant, resuspend the beads in 20 uL of cold NPB, and proceed with downstream processing (RNA isolation or ChIP). Alternatively, nuclei/beads can be stored at -80°C until further use.
11. To view the purified nuclei under a microscope, add 1 uL of 0.2 ug/uL DAPI to each 25 uL sample (taken at step 7) and place on ice for 5 minutes. Count the number of nuclei using a hemocytometer (*see* **Notes 7-9**) (**Fig. 2**).

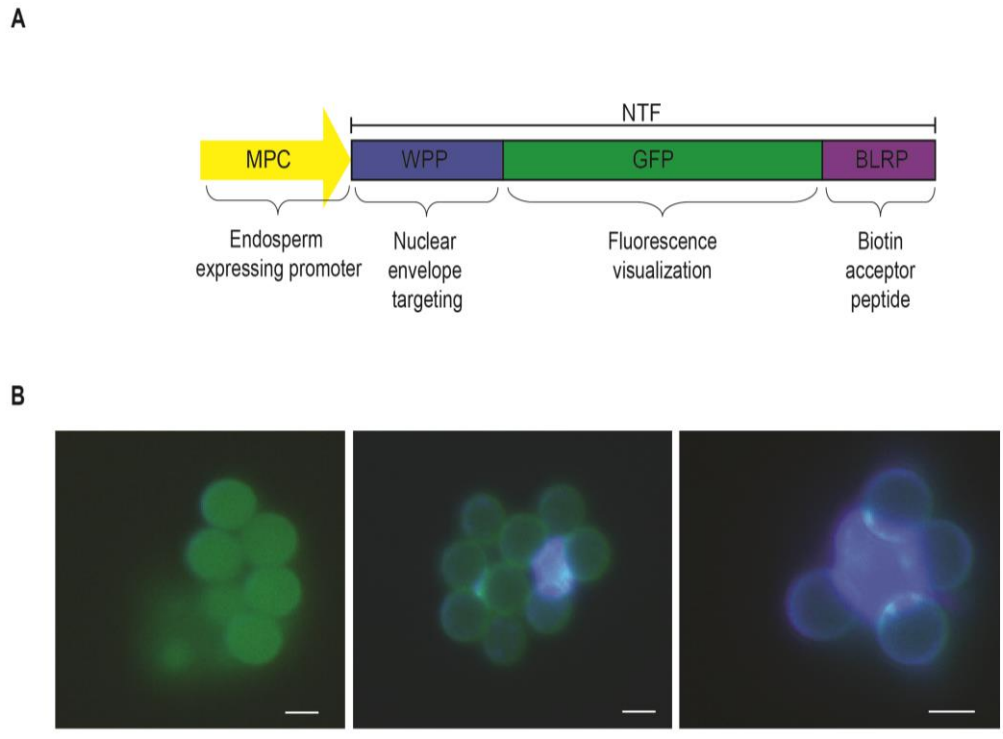


Fig. 2 Purification of tagged nuclei using the isolation of nuclei tagged in specific cell types (INTACT) system **(10)**. **a.** Partial cassette of the transgenic vector showing the three-part structure of nuclear targeting fusion protein (NTF) is shown. The chimeric protein consists

of the WPP domain of RanGAP1 for nuclear envelope targeting, GFP to allow visualization and the biotin ligase recognition peptide (BLRP), which is biotinylated by BirA. **b.** Fluorescence microscopy images of NTF-labeled nuclei that have been bound by streptavidin-coated magnetic beads prior to capture on a magnet. Biotinylated nuclei are shown in blue and GFP fluorescence from the beads themselves, is shown in green. Scale bar, ~3 μm in each panel. The right panel is at a higher magnification than the left and center panels.

3.3 RNA extraction, purification and DNase Treatment

1. Add 0.5 mL of TRIzol reagent to the beads from step 3.2 above and vortex immediately (*see* **Note 10**).
2. Transfer the solution to 1.5 mL tubes.
3. Centrifuge at 12,000g for 5 minutes at 2 to 4°C.
4. Remove the supernatant to a new 1.5 mL tube.
5. Add 200 μL of chloroform and shake vigorously by hand for approximately 15 seconds.
6. Let the tube stand at RT for 3 minutes.
7. Centrifuge the tube at 12,000g for 15 minutes at 4°C.
8. Carefully transfer the upper aqueous phase to a new 1.5 mL tube (ensure no interface debris are transferred) (*see* **Note 11**).
9. Add 0.5 mL of isopropyl alcohol and mix thoroughly.
10. Let the mix stand at room temperature for 10 minutes.
11. Centrifuge at 12,000g for 10 minutes at 4°C.
12. Carefully discard the supernatant (tip the tube with the pellet position angled up and away from you and pipet out the supernatant). The pellet may be slightly glassy and transparent or may not be visible at all.

13. Add 1 mL of 75% ethanol.
14. Vortex briefly and centrifuge at 12,000g for 5 minutes at 4°C.
15. Discard the ethanol and allow the pellet to air-dry for 10 minutes.
16. Dissolve the pellet in 20 µL of RNase-free water by very gently mixing with a pipette.
17. Check quantify and purity of the RNA by using a fluorimeter, RNA quality can be assessed by separation of 1µg on 2.5% agarose gel. Store the RNA at –80°C until further use.
18. Remove contaminating genomic DNA from the RNA by treating with the Turbo DNA-free kit in a 0.6-mL tube according to the manufacturer's instructions.
19. Remove the treated RNA to a new 0.6-mL tube.

3.4 Generation of RNA-Seq libraries

1. Remove ribosomal RNA (rRNA) from DNase treated RNA by treating with the Ribo-Zero™ magnetic kit according to the manufacturer's instructions.
2. Prepare sequencing libraries for the RNA from step 1 by using the NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina as per the manufacturer's instructions.
3. Check the prepared libraries for quality and quantity by separation on a Agilent High Sensitive DNA chip. Store the remaining library at -20°C until further use.

3.5 NextSeq sequencing

1. Prepare the sample sheet using the Illumina Experiment Manager by following the manufacturer's protocol (*see* **Note 12**).
2. Dilute the constructed libraries to 4 nM in resuspension buffer (RSB) based on the concentrations determined by the bioanalyzer. From this point onwards, keep the libraries on ice.
3. Dilute the PhiX control library to 4 nM by adding 15 µL of RSB to 10 µL of the 10 nM PhiX control library (*see* **Note 13**).
4. Denature the pooled libraries and PhiX control library separately by adding 5 µL 0.2 M NaOH to 5 µL of the 4 nM libraries (*see* **Note 14**).
5. Vortex thoroughly to mix and incubate at RT for 5 min. Add 5 µL 200 mM Tris-HCl, pH 7.0 and vortex briefly before centrifuge at 280g for 1 minute.
6. Dilute the denatured pooled libraries and PhiX control library separately to 20 pM by adding 985 µL pre-chilled HT1 to 15 µL denatured libraries.
7. Dilute the 20 pM pooled libraries and PhiX control library separately to 1.8 pM by adding 1183 µL pre-chilled HT1 to 117 µL 20 pM denatured libraries.
8. Combine 13 µL of the 1.8 pM PhiX control library with 1287 µL of the 1.8 pM pooled libraries and vortex to mix.
9. Load 1300 µL of the final sample into the cartridge. Ensure that any air bubbles are removed by gently tapping the cartridge.
10. Perform the sequencing run according to the manufacturer's protocol.

3.6 Bioinformatics analysis to identify lncRNAs

1. Trim adaptors and low-quality sequences from raw reads by using trim_galore with following parameter: -- stringency 6.

2. Align the trimmed reads against the *Arabidopsis thaliana* genome TAIR10 assembly by using TopHat2 with the following parameters:

   -N 5 -- read-edit- dist 5.

3. Merge the aligned reads from all samples by using SAMtools and assembled the reads into transcripts by using cufflinks by using the parameter:

   -- library-type fr-firststrand -u.

4. Remove transcripts shorter than 200 nt by a custom script.
5. Determine the genomic locations of long transcripts from step 4 by comparing the genomic coordinates against the reference genes of TAIR10 and annotate the transcripts into either; gene, intergenic, intronic or antisense.
6. Determine the protein-coding potential of the annotated intergenic, intronic and antisense transcripts by undertaking the following two steps: 1) Sequence similarity search against the SWISS-PROT protein database; and 2) Predict Open Reading Frame(s) (ORF). Transcripts that have no sequence similarity to proteins in SWISS-PROT or no ORFs longer than 20 amino acids are candidate lncRNAs.
7. Save and export the sequences of novel lncRNAs.

3.7 Functional analysis of lncRNAs

1. Amplify target lncRNAs by using PCR from either genomic DNA or cDNA.
2. To overexpress or knockdown target lncRNAs by using dsRNA continue with the following steps (step 3). Alternatively, to strand specifically

knockdown a lncRNA follow the procedures in step 4. Then proceed to step 5, transformation using *Agrobacterium tumefaciens*.

3. To overexpress or knockdown target lncRNAs by using dsRNA: Clone the PCR amplicons into pENTR™/D donor vector as per the manufacturer's instructions. Carry out a LR reaction between the recombinant pENTR™/D donor vector and destination binary vector (vector pLEELA for overexpression or pJahwol18 for knockdown) as per the manufacturer's instructions.
4. To strand specifically knockdown a lncRNA: Strand specific knockdown of a lncRNA may be desirable, for example if it is antisense to another transcript, and can be performed by using artificial miRNAs (amiRNAs). Generate, candidate amiRNA sequences that target your lncRNA by using the web app for the automated design of artificial microRNAs, WMD3 (http://wmd3.weigelworld.org/cgi-bin/webapp.cgi). From the list of candidate amiRNA sequences, select an amiRNA sequence, include 5' and 3' miRNA319 stem loop and amiRNA* sequences and order the sequence as a gBLOCK® from IDT. Clone the amiRNA gBLOCK® into vector pRS300 to generate 35S:amiRNA. Digest the vector with *Pvu*II to remove the 35S:amiRNA cassette and clone the cassette into binary vector pGreenII.
5. Introduce the recombinant binary vectors into *Agrobacterium tumefaciens* strain AGL 1 by electroporation. Select transformants by using the appropriate antibiotics (Rifampicin and vector conferred antibiotic resistance, e.g. Ampicillin) (*see* **Note 15**).

6. Grow 20 plants under long day conditions for 4 weeks until the inflorescence is approximately 10 cm long. Transform the binary vector into *Arabidopsis thaliana* plants by using the floral dipping method.
7. Grow a 450 mL culture of Agrobacterium containing the binary vector overnight (*see* **Note 16**).
8. Repeatedly dip the flowering plants into the Agrobacterium solution for 60 seconds. Place the dipped plants horizontally onto a tray such that the Agrobacterium solution does not run down onto the leaves. Cover the tray by using a plastic cover and place into a growth chamber for 24 hours.
9. Remove the plastic cover and place the dipped plants vertically.
10. Collect seeds approximately 3 weeks later.
11. Plant on soil ~2,000 dried seeds and 1 week after germination, select for transformed plants by an aerial application of the herbicide BASTA®. Apply the herbicide to run-off point. Repeat BASTA® application 1 week later (*see* **Note 17**). Identify approximately 20 T_1 plants.
12. From mature T_1 plants, harvest seed, store in a cool, low humidity place and dry the seed for 2 weeks.
13. Plant approximately 100 seeds for each of the 20 T_1 lines. One week later, count the number of germinated T_2 seedlings.
14. Spray BASTA® onto the T_2 seedlings and 5 days later count the number of resistant seedlings. Select T_2 lines that have 3:1 (resistant: sensitive) segregation, and grow 20 plants and harvest the seed.
15. Perform phenotypic and molecular analysis on T_3 plants after selection with BASTA® (**Fig. 3**). Homozygous transgenic plants can be identified by progeny screening after BASTA® application.

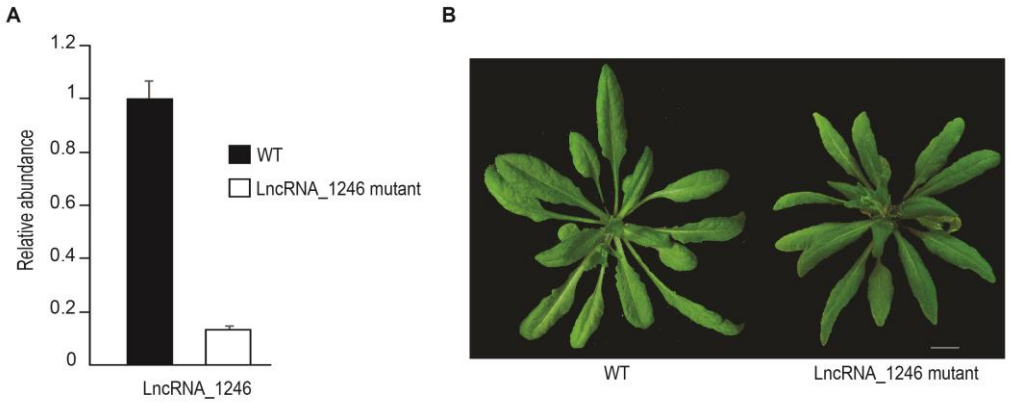


Fig. 3 Mutation of *LNCRNA_1246* caused smaller rosette leaves than wild type (WT) **a.** RT_PCR for the expression of *LNCRNA_1246* in the mutants and WT. Mean ± SD are shown. **b.** Images of *LNCRNA_1246* mutant and WT at flowering stage. Scale bar is 1 cm.

3.8 Chromatin isolation by RNA purification (ChIRP)

1. Sow 1g of lncRNA mutant and wild type control seeds on ½ Murashige and Skoog (MS), 1% sucrose agar plates.
2. Seven days after germination, crosslink seedlings with 1% formaldehyde in 1X PBS solution by using vacuum infiltration at 4ºC for 15 minutes (*see* **Note 18**).
3. Terminate crosslinking by adding 0.125M glycine and vacuum infiltrate for another 5 minutes.
4. Wash the cross-linked seedlings three times with cold, sterile water.
5. Grind the seedlings to a fine powder in a mortar and pestle that has been pre-chilled with liquid nitrogen. Add 1 mL of ground powder to 10 mL of buffer 1.
6. Keep the samples on ice for 10 min. Filter the samples through two 40 µM strainers and then centrifuge at 2,000 g at 4ºC for 30 minutes.

7. Remove the supernatant and resuspend the pellet in 1 mL of buffer 2, centrifuged at 15,000 g at 4ºC for 10 minutes.
8. Remove the supernatant and resuspended the pellet in 300 µL of buffer 3, layer on top another 300 µL of fresh buffer 3, centrifuged at 15,000 g at 4ºC for 1 hour.
9. Remove the supernatant and re-suspended the pellets in 300 µL of lysis buffer, sonicated (15s ON/60s OFF) until the DNA is fragmented into 200±500 bp length fragments.
10. Centrifugation at 15,000 g at 4ºC for 15 minutes. Dilute the chromatin in the supernatant with 2 volumes of hybridization buffer.
11. Add the Stellaris® designed biotin labelled lncRNA probes to 150 µL of diluted chromatin to give a final concentration of 100 pmol/uL, and incubate by end-to-end rotation at 4ºC overnight.
12. For each sample, wash 50 µL of streptavidin-magnetic C1 beads with 6 volumes of lysis buffer, and repeat three times. Block the beads by adding 500 ng/µL of yeast total RNA and 1 mg/mL BSA and incubate for 1 hour at RT, then washed again in lysis buffer before re-suspending in the original volume of 50 uL.
13. Mix the samples from step 11 and the washed beads from step 12 at 4ºC for 2 hours
14. Wash the captured beads five times with wash buffer. The beads are now ready for different elution protocols depending on the downstream assays.
15. RNA elution: Resuspend the beads in 10x the original volume of RNA elution buffer and boil for 15 minutes. Purify the RNA by following the RNeasy mini column procedure according to the manufacturer's protocol.

Detect the enriched transcripts by quantitative reverse-transcription PCR (qRT-PCR).

16. DNA elution: Resuspended the beads in 3x the original volume of DNA elution buffer with a cocktail of 100 µg/mL RNase A and 0.1 U/µL RNase H at 37^{o}C with end-to-end rotation. Reverse-crosslink the chromatin by incubation at 65^{o}C overnight. Purify the DNA by phenol: chloroform: isoamyl-alcohol extraction. Measure the enriched DNA by qPCR or high-throughput sequencing relative to the negative control.
17. DNA elution: resuspend the beads in 3x original volume of DNase buffer (100 mM NaCl and 0.1% NP-40), and elute the protein with a cocktail of 100 µg/mL RNase A and 0.1 U/µL RNase H, and 0.1 U/µL DNase I and incubate at 37^{o}C for 30 minutes.
18. To observe enriched proteins, add 0.2 volume of 5x laemmeli buffer to the sample and negative control, boil for 5 minutes and then separate the samples on an acrylamide gel. Silver stain the gel to observe the RNA-binding proteins that are present only in the sample and not negative control.

4 Notes

1. If using Eppendorf tubes, prepare the tubes with a small hole in the lid to prevent the tube opening as the tube warms due to residual liquid nitrogen expanding.
2. For storage in RNA*later*^{TM} solution, add approximately 3 × the volume of RNA*later*^{TM} : 1 × tissue.

3. Tissue stored in RNA*later*™ and frozen (−80°C) was defrosted just enough to remove the tissue from the RNA*later*™ solution prior to extraction. Repeated removal from storage often leads to reduced quality of RNA.
4. It is very important to minimize the time between removal of the tissue from the plant and immersion in liquid nitrogen.
5. The amount of tissue required to achieve acceptable yields of RNA varies according to the material. Tissues with a high-water content require larger amounts of tissue for the same yield of RNA.
6. It is essential to grind the tissue as finely as possible and maintain the samples as cold as possible during grinding to avoid RNA degradation.
7. The actual yield of purified nuclei is generally around 50% of the theoretical yield for the cell types we have examined. Therefore, it is recommended to begin with an amount of tissue that will yield at least double the required number of nuclei. The number of nuclei should yield 100-200 ng of total RNA when purified by using the RNeasy Micro kit.
8. For crosslinked chromatin immunoprecipitation experiments, starting tissue can be treated with formaldehyde, quenched, washed, and used directly in the above protocol without any alterations.
9. Use of this protocol with other types or amounts of tissue may require optimization. The most important parameters seem to be the number of beads used per mass of tissue and the volume of solution when capturing the beads after nuclei binding. For larger scale purifications, the DynaMag-50 magnet can be used to capture beads in volumes up to 40 mL.

10. If the beads were stored at -80ºC, do not let them thaw without being in the presence of extraction buffer. It is important to ensure that the beads do not form a clump, where the outside of the clump is in contact with the buffer, but the inner beads are not.
11. Plant tissues also contain other compounds that interfere with RNA extraction (such as polysaccharides, lipids, proteins). If these compounds are not removed in the first steps (discarded in the aqueous phase), they will remain through the rest of the extraction. Therefore, it is very important not to remove any debris or interface material during the chloroform extraction.
12. The sample sheet is required to insert the sample names and adaptor indices used for each sample. We have selected the “Other” as the category followed by “FASTQ only”. This option generates FASTQ files only and also enables the deselection of down-stream processing steps like adaptor trimming, allowing trimming and mapping to be performed separately.
13. The prepared PhiX library is added to the pooled amplicon libraries as an internal control for the MiSeq sequencing run.
14. It is best to prepare fresh 0.2 M NaOH for the denaturation of libraries.
15. Other binary vectors may confer different antibiotic resistance.
16. Do not cover the plants for more than 24 hours. Excess humidity over a long time can yield low number of transformed plants.
17. Resistant seedlings should be screened for the expression of target lncRNAs by RT-PCR using specific primers. The highest or lowest

expression of target lncRNAs for overexpression or knockdown experiments should be used for further analysis, respectively.

18. When the solution starts to boil, stop and slowly release the vacuum. Repeat the vacuum infiltration until the seedlings sink to the bottom after the vacuum is released. Do not exceed 15 minutes total exposure time to the formaldehyde.

Acknowledgements

This work was supported by an Australian Research Council Future Fellowship (FT130100525) awarded to IS and a MOET-VIED PhD scholarship awarded to TD.

References

1. Chekanova JA, Gregory BD, Reverdatto SV, Chen H, Kumar R, Hooker T, Yazaki J, Li P, Skiba N, Peng Q, Alonso J, Brukhin V, Grossniklaus U, Ecker JR, Belostotsky DA (2007) Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the *Arabidopsis* transcriptome. Cell 131(7):1340–1353.
2. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316(5830):1484–1488.

3. Burgess AL, David R, Searle IR (2015) Conservation of tRNA and rRNA 5-methylcytosine in the kingdom Plantae. BMC Plant Biol. 15:199.
4. Wang D, Qu Z, Yang L, Zhang Q, Liu ZH, Do T, Adelson DL, Wang ZY, Searle I, Zhu JK (2017) Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in Plants. Plant J 90:133-146.
5. Ashby R, Forêt S, Searle I, Maleszka R (2016) MicroRNAs in Honey Bee Caste Determination. Scientific Reports 6:18794.
6. Jin J, Liu J, Wang H, Wong L, Chua N-H (2013) PLncDB: plant long noncoding RNA database. Bioinformatics 29(8):1068–1071.
7. David R, Burgess A, Parker, B, Li, J, Pulsford, K, Sibbritt, T, Preiss, T, Searle, I (2017) Transcriptome-wide mapping of RNA 5-methylcytosine in *Arabidopsis* mRNAs and ncRNAs. The Plant Cell 29(3) 445-460.
8. Wang H, Chung PJ, Liu J, Jang I-C, Kean MJ, Xu J, Chua N-H (2014) Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in *Arabidopsis*. Genome Res 24(3):444–453.
9. Zhang Y-C, Liao J-Y, Li Z-Y, Yu Y, Zhang J-P, Li Q-F, Qu L-H, Shu W-S, Chen Y-Q (2014) Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. Genome Biol 15(12):512.
10. Wang HLV and Chekanova JA (2017) Long Noncoding RNAs in Plants. In: Rao M. (eds) Long Non Coding RNA Biology. Advances in Experimental Medicine and Biology, vol 1008. Springer, Singapore.

11. Deal RB, Henikoff S (2011) The INTACT method for cell type–specific gene expression and chromatin profiling in *Arabidopsis thaliana*. Nature Protocol 6(1):56-68.
12. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349.
13. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453:1239–1243.
14. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5:621–628.
15. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133:523–536.
16. Cloonan N, et al (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods 5:613–619.
17. Marioni J, Mason C, Mane S, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18(9):1509-1517
18. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M (2008) Profiling the HeLa S3 transcriptome using randomLy primed cDNA and massively parallel short-read sequencing. Biotechniques 45:81–94.

19. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al (2012) Landscape of transcription in human cells, Nature 489: 101–108.
20. Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua N-H (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. Plant Cell 24(11):4333–4345.
21. Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, et al (2014) Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat. Genet. 46: 558–566.
22. Guil S, Soler M, Portela A, Carrère J, Fonalleras E., Gómez A, et al (2012) Intronic RNAs mediate EZH2 regulation of epigenetic targets. Nat. Struct. Mol. Biol. 19:664–670.
23. Qiu J, Wang Y, Ding J, Jin H, Yang D, Hua K (2015) The long noncoding RNA HOTAIR promotes the proliferation of serous ovarian cancer cells through the regulation of cell cycle arrest and apoptosis. Exp. Cell Res. 333(2):238-48.
24. SheikMohamed J, Gaughwin PM, Lim B, Robson P, Lipovich L (2010) Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. RNA 16: 324–337.
25. Liu SJ, Horlbeck, Cho SW, Birk HS, Malatesta M, He D, Attenello FJ, Villalta JE, Cho MY, Chen Y, Mandegar MA, Olvera MP, Gilbert LA, Conklin BR, Chang HY, Weissman JS, Lim DA (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science 355(6320):eaah7111.

Chapter 3: Quantitative and Single Nucleotide Resolution Profiling of RNA 5-methylcytosine

Jun Li[1], Xingyu Wu[1], Trung Do[1], Vy Nguyen[1], Jing Zhao[1], Pei Qin Ng[1], Alice Burgess[1], Rakesh David[1] and Iain Searle[1*]

[1]School of Biological Sciences, Department of Molecular and Biomedical Sciences, The University of Adelaide, Australia

[*]Corresponding Author: iain.searle@adelaide.edu.au

Submitted for a chapter in the Methods in Molecular Biology: Epitranscriptomics.

Statement of Authorship

Title of paper	Quantitative and single nucleotide resolution profiling of RNA 5-methylcytosine
Publication Status	Accepted
Publication details	In press Spinger Science + Business, Media, LLC, New York

Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. The candidate's stated contribution to the publication is accurate (as detailed below)
ii. Permission is granted for the candidate in include the publication in the thesis and
iii. The sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Candidate	Trung Q. Do		
Contribution to chapter	Composed figure 3. Edited the manuscript.		
Overall percentage (%)	10%		
Signature		Date	15/3/2018

Name of Co-Author	Jun Li		
Contribution to chapter	Wrote and edited the manuscript. Prepared figures. Overall contribution 40%.		
Signature		Date	16/03/2018

Name of Co-Author	Xingyu Wu		
Contribution to chapter	Edited the manuscript. Overall contribution 10%.		
Signature		Date	16/03/2018

Name of Co-Author	Vy Nguyen		
Contribution to chapter	Edited the manuscript. Overall contribution 5%.		
Signature		Date	16/3/18

Name of Co-Author	Jing Zhao		
Contribution to chapter	Edited the manuscript. Overall contribution 5%.		
Signature		Date	16/3/18

Name of Co-Author	Pei Qin Ng		
Contribution to chapter	Edited the manuscript. Overall contribution 5%.		
Signature		Date	21/3/2018

Name of Co-Author	Alice Burgess		
Contribution to chapter	Edited the manuscript. Overall contribution 5%.		
Signature		Date	19/3/18

Name of Co-Author	Rakesh David		
Contribution to chapter	Edited the manuscript. Overall contribution 5%.		
Signature		Date	16/03/18

Name of Co-Author	Iain R. Searle		
Contribution to chapter	Supervised development of work and edited the manuscript. Overall contribution 15%.		
Signature		Date	16/3/18

Abstract

RNA has co-evolved with numerous post-transcriptional modifications to sculpt interactions with proteins and other molecules. One of these modifications is 5-methylcytosine (m^5C) and mapping the position and quantifying the level in different types of cellular RNAs and tissues is an important objective in the field of epitranscriptomics. Both in plants and animals bisulfite conversion has long been the gold standard for detection of m^5C in DNA but it can also be applied to RNA. Here, we detail methods for highly reproducible bisulfite treatment of RNA, efficient locus-specific PCR amplification, detection of candidate sites by sequencing on the Illumina MiSeq platform and bioinformatic calling of non-converted sites.

Key words Bisulfite conversion, Epitranscriptome, Fluidigm Access Array, Illumina, next-generation sequencing, 5-methylcytosine

1 Introduction

Cellular RNAs can be modified, or decorated, with more than one hundred and twenty chemically and structurally distinct nucleoside modifications [1]. The emerging field of epitranscriptomics [2] has been enabled by the development of high-throughput mapping methods for RNA modifications, typically based on second generation sequencing. Transcriptome-wide positions of N1-methyladensosine (m^1A, [3-5]), N6-methyladenosine (m^6A, [6,7]), 5-methylcytosine (m^5C, [8]) and pseudouridine [9] have each been reported in this way. To detect m^5C in RNA, a range of methods have been developed, including the indirect (aza-IP [10], miCLIP [11]) immunoprecipitation of methylated RNA or

direct methods (meRIP, [7]). Of particular interest here, the bisulfite conversion approach previously used for DNA has been adapted to RNA [12,13]. Bisulfite conversion of nucleic acids takes advantage of the differential chemical reactivity of m^5C compared to unmethylated cytosines; unmethylated cytosines are deaminated to uracil while m^5C remains as a cytosine.

The RNA bisulfite conversion method has been applied to animals and plants [8,14] using second generation sequencing, for example Illumina, based transcriptome-wide readout and mapped thousands of novel candidate m^5C sites in a diverse array of RNAs, including mRNAs and long noncoding RNAs (lncRNAs). Here, we detail protocols for RNA bisulfite conversion, locus-specific PCR amplification of up to 2,304 amplicons, and bioinformatics calling of converted or non-converted sites. Sequencing of PCR amplicons is conveniently done on the Illumina MiSeq, as this affords multiplexing of multiple distinct amplicons while still achieving ample read depth for estimating the proportion of m^5C at targeted positions. For instance, each of the 96 Fluidigm indexed adaptors could be assigned to a separate RNA derived from different tissues, and 96 multiple PCR amplicons per sample could be included in the sequencing pool, potentially generating thousands of independent quantitative measurements of the m^5C levels in a single MiSeq run (Fig. 1).

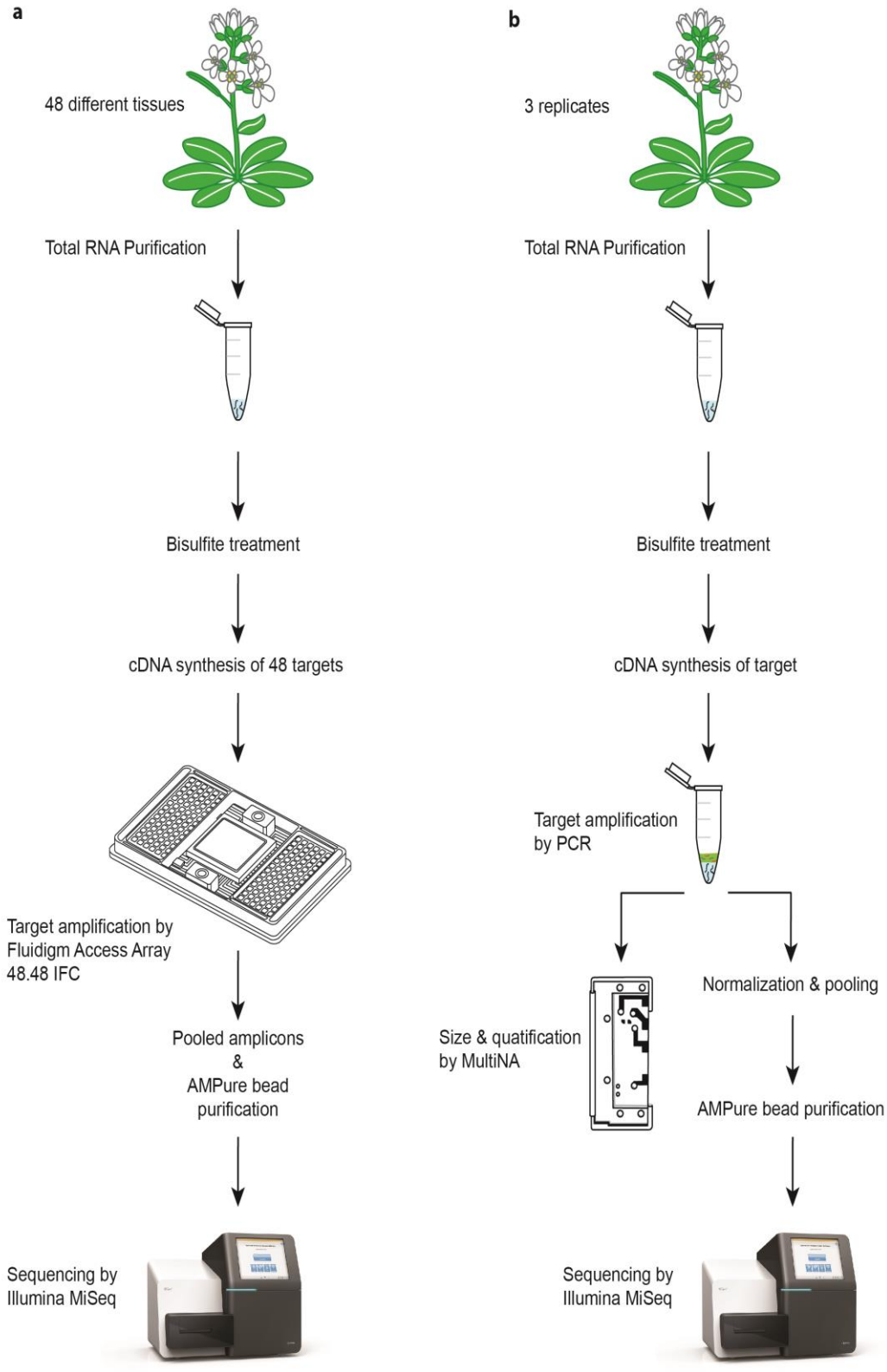


Fig. 1 Protocol overview showing the workflow for either parallel or single amplicon amplification for effective detection of m^5C. (**a**) Parallel amplification and sequencing of up to 2304 amplicons across 48 tissues and 48 primer pairs. Forty-eight different tissues can be selected, total RNA

isolated and purified, spiked with MGFP in vitro transcribed control RNA and bisulphite converted. Bisulphite converted RNA is reverse transcribed (RT) to cDNA using gene specific RT primers that includes the positive control *MAG5* (AT5G47480) and negative control MGFP. Target regions are PCR amplified using a Fluidigm Access Array Integrated Fluidic Circuit (IFC), and up to 2,304 amplicons are harvested and eluted pools quantified. Equal concentrations of the pools are combined into a final pool, purified using AMPure beads, accurately quantified, PhiX control library spiked-in and subjected to sequencing on the Illumina MiSeq platform. (**b**) Single amplicon amplification and sequencing. A single tissue is selected, RNA isolated and purified in triplicate, spiked with MGFP *in vitro* transcribed control RNA and bisulphite converted. Bisulphite conversion and cDNA synthesis is the same as outlined above except that a specific target RT primer is used. The target amplicon is PCR amplified, triplicate amplicons are pooled, size and concentration is assessed on a Shimadzu MultiNA and amplicons pooled at equal concentration. Pooled amplicons are purified, PhiX control library spiked-in and subjected to sequencing on the Illumina MiSeq platform.

2 Materials

Prepare all solutions using RNase-free and DNase-free H_2O and analytical grade reagents. Store and prepare all reagents at room temperature unless indicated otherwise. Prepare and perform bisulfite conversion, cDNA synthesis, and PCR amplification experiments in an RNase-free area. Follow all state or national safety and waste disposal regulations when performing experiments.

2.1 *In vitro* Transcription Components

1. phMGFP Monster Green® Fluorescent reporter vectors (Promega)
2. HiScribe T7 Kit (NEB)
3. TURBO™ DNase (ThermoFisher Scientific)
4. Phase Lock Gel™ QuantBio (2.0 mL) (VWR)

5. UltraPure™ Phenol:Water (3.75:1 v/v) (ThermoFisher Scientific)
6. Chloroform
7. Glycogen (5 mg/mL) (ThermoFisher Scientific)
8. Agilent RNA 6000 Nano Kit (Agilent)

2.2 Sodium Bisulphite conversion components

1. Sodium bisulfite solution: 40 % (w/v) sodium metabisulfite (Merck), 0.6 mM hydroquinone, final pH 5.1 (Merck).

   To prepare the sodium bisulfite solution, prepare the following:

   0.6 M Hydroquinone: Weigh 66 mg hydroquinone and place into a 1.5 mL tube. Add H_2O to 1 mL and cover in foil to protect from light! Place in an orbital shaker to dissolve.

   40 % (w/v) sodium bisulfite: Dissolve 4 g sodium metabisulfite in 10 mL H_2O in a 50 mL falcon tube and vortex until it completely dissolves. Add 10 µL 0.6 M hydroquinone to the 40 % sodium bisulfite solution, vortex, and adjust pH to 5.1 with 10 M NaOH. Filter the solution through a 0.2 µm filter. Cover in foil to protect from light (*see* **Note 1**).
2. 1 M Tris–HCl, pH 9.0
3. Micro Bio-Spin. P-6 Gel Columns, Tris buffer (Bio-Rad)
4. Mineral oil
5. 75 % ethanol
6. 100 % ethanol
7. 3 M sodium acetate, pH 5.2
8. 5 mg/mL glycogen (ThermoFisher Scientific)

2.3 cDNA synthesis components

1. SuperScript III Reverse Transcriptase Kit (ThermoFisher Scientific)
2. 10 mM mixed dNTPs (Roche)
3. Single target priming- 20 µM gene specific oligo for each amplicon
4. Pool targets priming- 48 primers at 20 µM each

2.4 PCR amplicon amplification components

1. KAPA Biosystems HiFi DNA polymerase (KAPA Biosystems)
2. 10 mM mixed dNTPs (Roche)
3. T0.1E (10 mM Tris, pH 8.0, 0.1 mM EDTA)
4. Fluidigm Access Array Integrated Fluidic Circuit (IFC) 48.48 (Fluidigm)
5. FastStart High Fidelity PCR System, dNTPack (Roche)
6. 20X Access Array Loading Reagent (Fluidigm)
7. 1X Access Array Harvest Solution (Fluidigm)
8. 1X Access Array Hydration Reagent v2 (Fluidigm)
9. Access Array Barcode primers for Illumina Sequencers-384: Single Direction (Fluidigm)

2.5 MultiNA Microelectrophoresis System

1. DNA-500 Kit (Shimadzu)

2.6 PCR amplicon purification and Quantification

1. Agencourt® AMPure® XP beads (Beckman Coulter)
2. Library Quantification Kit (Universal) from KAPA Biosystems

2.7 Library sequencing components

1. 0.2 M NaOH
2. Illumina MiSeq Reagent Kit v3 (150 or 600 cycles) (*see* **Note 2**)

3 Methods

Carry out all procedures described below at room temperature unless otherwise stated.

3.1 RNA extraction, purification and DNase Treatment

Total RNA is extracted and purified directly from tissue with 1 mL of TRIzol® as per the manufacturer's protocol. RNA is then treated with TURBO™ DNase as per the manufacturer's protocol. Assess the integrity of the RNA by using a RNA 6000 Nano Chip on the Agilent. 2100 Bioanalyzer according to the manufacturer's protocol.

3.2 Generation of the MGFP *In Vitro* Transcript Spike-In Control

1. Linearise the phMGFP vector by using the restriction enzyme *XbaI* and purify the linearized DNA vector according to the HiScribe T7 kit protocol.
2. Perform *in vitro* transcription according to the HiScribe T7 kit protocol by using 1 µg of linearized DNA. An incubation period of 4 hr at 37^{o}C with the kit components is sufficient.
3. Add 2 U TURBO™ DNase and incubate at 37 ^{o}C for 30 min.
4. Transfer the reaction to a Phase Lock Gel™ tube and make the volume of the reaction up to 100 µL with ultrapure H_2O.

5. Add an equal volume of Phenol:Water and chloroform, shake vigorously for 15 s, and centrifuge at 15,000 g for 5 min.
6. Add the same volume of chloroform as in step 5 to the tube, shake vigorously for 15 s, and centrifuge at 15,000 g for 5 min again.
7. Transfer the aqueous phase to a clean 1.5 mL tube. Add 1/10 volume 3 M sodium acetate, 3 volumes of 100 % ethanol, and 1 µL glycogen, vortex, and precipitate the RNA overnight at −80 ^{o}C.
8. Centrifuge RNA at 17,000 g at 4 ^{o}C for 60 min and carefully remove the supernatant.
9. Add 1 mL 75 % ethanol to the RNA, invert 5 times and centrifuge at 7500 g at 4 ^{o}C for 10 min (*see* **Note 3**).
10. Carefully remove the supernatant and let the pellet air-dry for approximately 15 mins (*see* **Note 4**).
11. Resuspend the RNA in 25 µL of ultrapure H_2O.
12. Optional Step- Treat 5 µg of *in vitro* transcribed MGFP transcript with 2 U TURBO™ DNase according to the manufacturer's protocol at 37 ^{o}C for 30 min.
13. Assess the integrity and size of the MGFP *in vitro* transcripts by using an RNA 6000 Nano Chip on the Agilent 2100 Bioanalyzer according to the manufacturer's protocol (*see* **Note 5**).

3.3 Bisulphite Conversion of RNA

1. Add 1/2000 of the MGFP RNA transcript to 2 µg DNase treated purified total RNA. Increase the volume of the RNA sample to 20 µL with ultrapure H_2O.

2. Denature RNA by heating to 75 ^{o}C for 5 min in a heat block.
3. Preheat the sodium bisulfite solution to 75 ^{o}C, add 100 µL to the RNA, vortex thoroughly, and briefly, 13K g for 1min, spin in a microcentrifuge.
4. Overlay the reaction mixture with 100 µL of mineral oil. Cover the tube in aluminium foil to protect the reaction mixture from light (*see* **Note 6**).
5. Incubate at 75 ^{o}C for 4 hr in a heat block.
6. About 15 min before the bisulfite conversion reaction is complete, prepare two Micro Bio-Spin Columns for each conversion reaction by allowing the Tris solution in the column to drain into a collection tube. Discard the Tris flow-through, place the column back into the collection tube, and centrifuge at 1000 g for 2 min. Transfer each column to a clean 1.5 mL tube (*see* **Note 7**).
7. Remove the bisulfite reaction mixture from the heat block and gently transfer the aqueous layer (that is under the mineral oil) containing the sodium bisulphite/RNA mixture to the Micro Bio-Spin column (*see* **Note 8**).
8. Centrifuge at 1000 g for 4 min.
9. Carefully transfer the eluate into the second Micro Bio-Spin column placed in a 1.5 mL tube and repeat step 8.
10. Preheat the temperature of the heat block to 75 ^{o}C in preparation for step 12.
11. Add an equal volume of 1 M Tris–HCl (pH 9.0) to the second eluate, vortex, spin briefly, and then overlay with 175 µL of mineral oil. Cover the tube in aluminium foil to protect the reaction mixture from light.
12. Incubate at 75 ^{o}C for 1 hr in the heat block.

13. Transfer the bottom aqueous layer containing the RNA to a clean 1.5 mL tube.
14. Precipitate the bisulphite treated RNA by following steps 7-11 in section 3.2, and resuspend the bisulfite-converted RNA in H_2O (*see* **Note 9**).

3.4 Bisulphite oligonucleotide primer design for cDNA synthesis and PCR

1. For efficient parallel amplification of 48 target amplicons on the Fluidigm Access Array, use targeted cDNA synthesis to reduce amplification of spurious amplicons. Targeted cDNA synthesis is achieved by designing reverse transcriptase (RT) primers 30-40 nt 3' of the cytosine(s) to be assayed. N.B. Design the RT primers such that they avoid areas of bisulphite-converted cytosines as inefficient BS conversion may result in unconverted cytosines and biasing later amplification. See Figure 2.

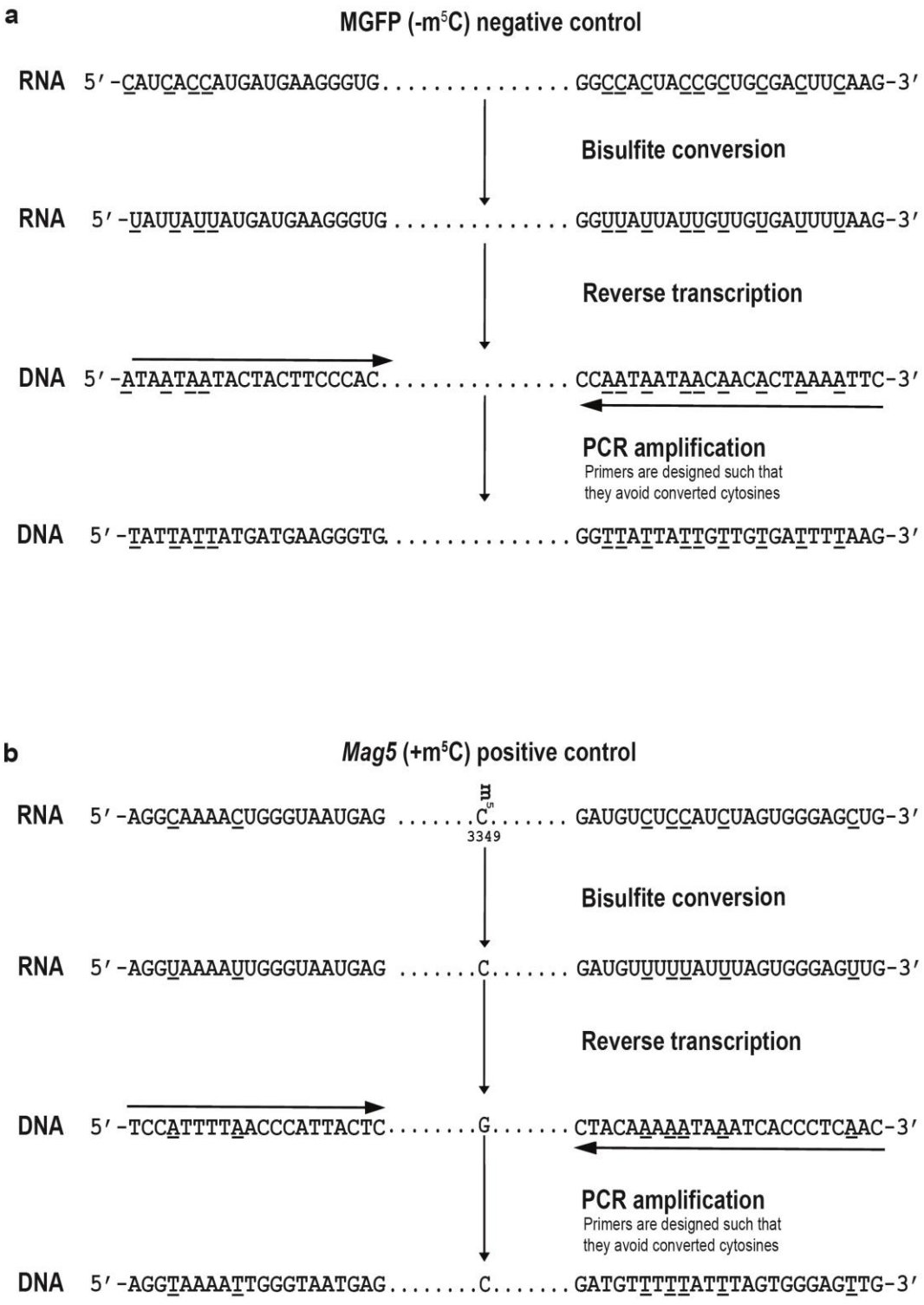


Fig 2. Overview of bisulphite conversion of RNA, reverse transcription to cDNA and PCR amplification. (**a**) In the *in vitro* transcribed MGFP sequence, unmodified cytosines (underlined) are converted to uracil, reverse transcribed (RT) by reverse transcriptase to cDNA and then PCR amplified. RT and PCR primers are designed to avoid stretches of

converted cytosines to prevent preferential amplification of converted sequences which may incorrectly indicate efficient bisulfite conversion. (**b**) In *MAG5* control and other candidate sequences, primers are designed to span areas containing converted cytosines to preferentially amplify converted sequences. C3349 is methylated in *Arabidopsis thaliana* and serves as a over-conversion control. Flanking cytosines are not methylated and should be completly converted. Primers are designed with a T_m of 59-61^{o}C, preferably with a 3' G nucleotide and to amplify PCR products of 170-200 bp.

2. Design primers for the first round of PCR amplification so that small amplicons are 170-200bp, to allow efficient amplification (*see* **Notes 10 and 11**). As the G/C content in the template is low, design long primers to ensure a T_m is in the rage of 59-61^{o}C. Add the CS1 sequence to the forward primer Gene Specific Sequence (GSS) and CS2 to the reverse primer GSS. For the second PCR amplification, use the forward primer containing the complementary sequences to the P5 Illumina flow cell combined and CS1 (P5_CS1) and the reverse primer containing the barcode, and complementary sequences to the P7 Illumina flow cell combined with CS2 (P7_BC_CS2) primer. (*see* **Note 12**). See Figure 3.

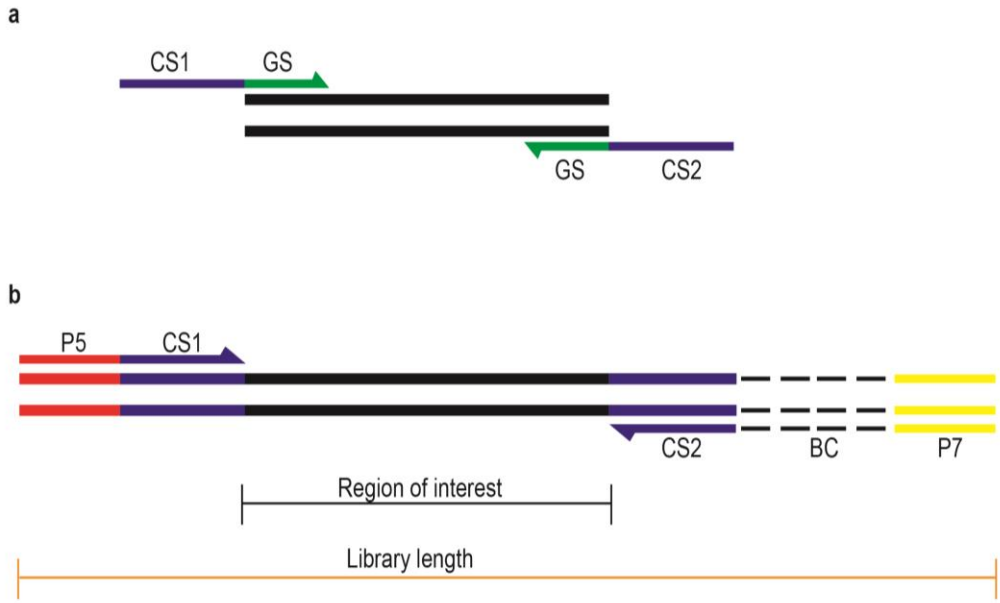


Fig 3. Overview of first and second PCR amplification of target regions. (**a**) For the first PCR, the forward PCR primer is designed with the gene specific sequence (GS) and universal forward tag called Common Sequence, CS1 (5'-TACGGTAGCAGAGACTTGGTCT-3') and reverse PCR primer is designed with the Gene Specific Sequence (GSS) and universal reverse tag called Common Sequence CS2 5'-ACACTGACGACATGGTTCTACA-3'). (**b**) For the second PCR, the forward primer is designed with the CS1 and Illumina P5 sequences and the reverse primer contains the CS2, barcoding and Illumina P7 sequences. The Fluidigm barcodes or indexes are 10 nt in length.

3.5 cDNA synthesis

1. Mix 500 ng of bisulfite-converted RNA, 1 µL of 1 mM dNTP mix, 2 µL of 10X pooled primer mix and add ultrapure H_2O to a final volume of 13 µL. Incubate the mix at 65 ^oC for 5 min to denature the RNA.
2. Reverse transcribe the bisulfite-converted RNA using the manufacturer's protocol. Add either pooled 48 RT primers for parallel Access Array amplification or random hexamers for single PCR amplicons.

3. Suggested controls- Include RT minus controls for each sample as the PCR primers are not necessarily designed to span exon-exon junctions. In the controls, use 1 µL of H_2O instead of reverse transcriptase.
4. After the reaction is complete, dilute the cDNAs 1:10 in ultrapure H_2O for PCR amplification.

3.6.1 Individual PCR amplification, quantification and pooling

1. For a 10 µl PCR, add 0.2 µl KAPA HiFi DNA Polymerase, 2 µl of 5 X HiFi Fidelity buffer (with $MgCl_2$), 0.3 µl 10 mM dNTP, 0.4 µl 10 µM forward primer (CS1_GSS), 0.4 µl 10 µM reverse primer (CS2_GSS), 1 µl diluted cDNA and H_2O to a final volume of 10 µl. Perform PCR for each amplicon in triplicate.
2. Gently finger vortex, briefly centrifuge and place into a preheated thermal cycler.
3. Perform a two-step thermal cycling PCR program. See Table 1 for more details.

Table 1. Two-step thermal cycling conditions for the amplification of individual amplicons.

Stage	Temperature (°C)	Time (s)
Initial denaturation	98	15
Step I (x 10 cycles)		
Denaturation	94	10
Annealing	60	30
Extension	72	15
Step II (x 20 cycles)		
Denaturation	94	10

Annealing	55	30
Extension	72	15
Final Extension	72	60
Hold	4	Forever

4. Pool the triplicates and perform an AMPure bead clean-up at a ratio of 1.8:1 to remove unincorporated primers and primer dimers. Repeat this step (see Notes 13 and 14).
5. Assess PCR amplicon size and concentration after separation on a Shimadzu Microchip Electrophoresis System MCE®-202 MultiNA.
6. Normalise the concentration of each amplicon in the experiment by dilution with H_2O to a concentration in the range of 0.5-5 ng/μl.
7. Perform the barcoding and Illumina adapter addition PCR. In a 10 μl PCR, add 0.2 μl KAPA HiFi DNA Polymerase, 2 μl of 5 X HiFi Fidelity buffer (with MgCl2), 0.3 μl 10 mM dNTP, 1 μl 10 μM forward primer (P5_CS1), 1 μl 10 μM reverse primer (P7_CS2), 2 μl diluted PCR amplicon and H_2O to a final volume of 10 μl.
8. Gently finger vortex, briefly centrifuge and place into a preheated thermal cycler.
9. Perform a two-step thermal cycling PCR program. See Table 2 for more details.
10. Assess PCR amplicon size and concentration after separation on a Shimadzu Microchip Electrophoresis System MCE®-202 MultiNA.
11. Pool the amplicons in equimolar concentration and purify them using AMPure beads according to the manufacturer's protocol. Use a ratio of

beads to pooled amplicons of 0.9:1 to ensure binding of amplicons and not primer dimers or unincorporated primers.

12. First estimate the DNA concentration using a Qubit dsDNA Broad Range Assay Kit according to the manufacturer's protocol. Then accurately assess the DNA concentration by using KAPA Library Quantification Kit for Illumina® Platforms. Perform serial dilution of the pooled amplicons such that fall into the dynamic range of the assay of 5.5 – 0.000055 pg/µL.

Table 2. One-step thermal cycling conditions for the addition of barcodes and Illumina adapters.

Stage	Temperature (°C)	Time (s)
Initial denaturation	98	15
One Step (x12 cycles)		
Denaturation	94	10
Annealing	63	30
Extension	72	30
Final Extension	72	120
Hold	4	Forever

3.6.2 Parallel PCR amplification using a Fluidigm Access Array Integrated Fluidic Circuit (IFC)

1. Prime the Access Array according to the manufacture's protocol.
2. Pre-warm the 20X Access Array loading reagent to room temperature before use. Prepare the pooled 48-oligonucleotide primer mix by mixing 2.0 µL 50 µM CS1-GS forward, 2.0 µL 50 µM CS1-GS reverse, 5.0 µL 20 X Access Array loading reagent and 91 µL of H_2O to a final volume of 100 µL.

3. Finger vortex the mix and centrifuge to spin the contents to the bottom of the tube.
4. Prepare the sample pre-mix solution by mixing 30 µL 10X FastStart High Fidelity Reaction Buffer (without $MgCl_2$), 54 µL 25 mM $MgCl_2$, 15 µL DMSO, 6.0 µL 10 mM dNTP mix, 3.0 µL FastStart High Fidelity Enzyme Blend, 15.0 µL 20X Access Array Loading Reagent and 57 µL H_2O.
5. Finger vortex the mix and centrifuge to spin the contents to the bottom of the tube.
6. Prepare the sample mix solutions, 48 in total, in a 96 well plate. Mix 3.0 µL sample pre-mix, 1.0 µL cDNA, 1.0 µL Access Array Barcode library primers.
7. Thoroughly vortex the solutions for at least 30 seconds and then centrifuge to spin down the contents to the bottom of the plate. NB. Each well should receive a uniquely barcoded primer pair.
8. Load 4.0 µL of the primer solution and 4.0 µL of the sample mix solution into the primer and sample inlets of the Access Array by using an 8-channel pipette.
9. Load the Access Array into the Pre-PCR IFC Controller AX according to the manufacture's protocol.
10. Place the Access Array onto the FC1 Cycler and start thermal cycling by selecting the protocol AA 48x48 Standard v1. The thermal cycling conditions are presented in Table 3.

Table 3. Multi-step thermal cycling conditions for the Access Array.

Temperature (°C)	**Time (s)**	**Number of cycles**
50	120	1

70	1200	1
95	600	1
95 60 72	15 30 60	10
95 80 60 72	15 30 30 60	2
95 60 72	15 30 60	8
95 80 60 72	15 30 30 60	2
95 60 72	15 30 60	8
95 80 60 72	15 30 30 60	5

11. To harvest the PCR products from the Access Array follow the manufacture's protocol. Once the final step is completed, eject the Access Array.
12. Collect the harvested PCR products into a labelled PCR 96-well plate. Carefully transfer 10 µL of harvested PCR products from each of the sample inlets into columns 1-6 of the labelled 96-well plate by using an 8-channel pipette.
13. Assess PCR amplicon size and concentration after separation on a Shimadzu Microchip Electrophoresis System MCE®-202 MultiNA.
14. Pool the amplicons in equimolar concentration and purify them using AMPure beads according to the manufacturer's protocol. Use a ratio of beads to pooled amplicons of 0.9:1 to ensure binding of amplicons and not primer dimers or unincorporated primers (see **Note 14**).
15. First estimate the DNA concentration using a Qubit dsDNA Broad Range Assay Kit according to the manufacturer's protocol. Then accurately assess the DNA concentration by using KAPA Library Quantification Kit for Illumina® Platforms. Perform serial dilution of the pooled amplicons such that fall into the dynamic range of the assay of 5.5 – 0.000055 pg/µL.

3.7 MiSeq sequencing

1. Prepare the sample sheet using the Illumina Experiment Manager by following the manufacturer's protocol (*see* **Note 15**).
2. Dilute the library to 10 nM in EBT buffer based on the concentrations determined by the qPCR. From this point, keep the libraries on ice.

3. Dilute the PhiX control library to 2 nM by adding 8 µL EBT buffer to 2 µL of the 10 nM PhiX control library (*see* **Note 16**).
4. Denature the pooled libraries and PhiX control library separately by adding 10 µL 0.2 M NaOH to 10 µL of the 2 nM libraries (*see* **Note 17**).
5. Vortex thoroughly to mix and centrifuge at 1000 x g for 30 seconds. Incubate at room temperature for 5 min.
6. Dilute the denatured pooled libraries and PhiX control library separately to 20 pM by adding 980 µL pre-chilled HT1 to 20 µL denatured libraries.
7. Dilute the 20 pM pooled libraries and PhiX control library separately to 10 pM by adding 500 µL pre-chilled HT1 to 500 µL 20 pM libraries.
8. Combine 100 µL of the 10 pM PhiX control library with 900 µL of the 10 pM pooled libraries and vortex to mix (*see* **Note 18**).
9. Load 600 µL of the final sample into the cartridge. Ensure that air bubbles are removed by gently tapping the cartridge.
10. Perform the sequencing run according to the manufacturer's protocol.

3.8 Bioinformatics analysis of data

1. To trim the Illumina adaptor sequences that were incorporated into the amplicons to permit sequencing of the 150 bp paired-end reads, use Trimmomatic in palindromic mode [15].
2. Sequencing reads can be aligned with meRanTK by using Bowtie2 internally [16]. Assemble reference sequences for the alignment by using the segments of RNA interrogated by sequencing prior to bisulfite conversion.

3. Extract the methylation state of individual cytosines from bisulfite-read alignments by using meRanCall. The number of reads can be extracted from the aligned sequencing reads in order to determine read coverage at a given cytosine.
4. To call differentially methylated cytosines use meRanCompare. The number of reads can be extracted from the aligned sequencing reads in order to determine read coverage at a given cytosine (Fig. 4).

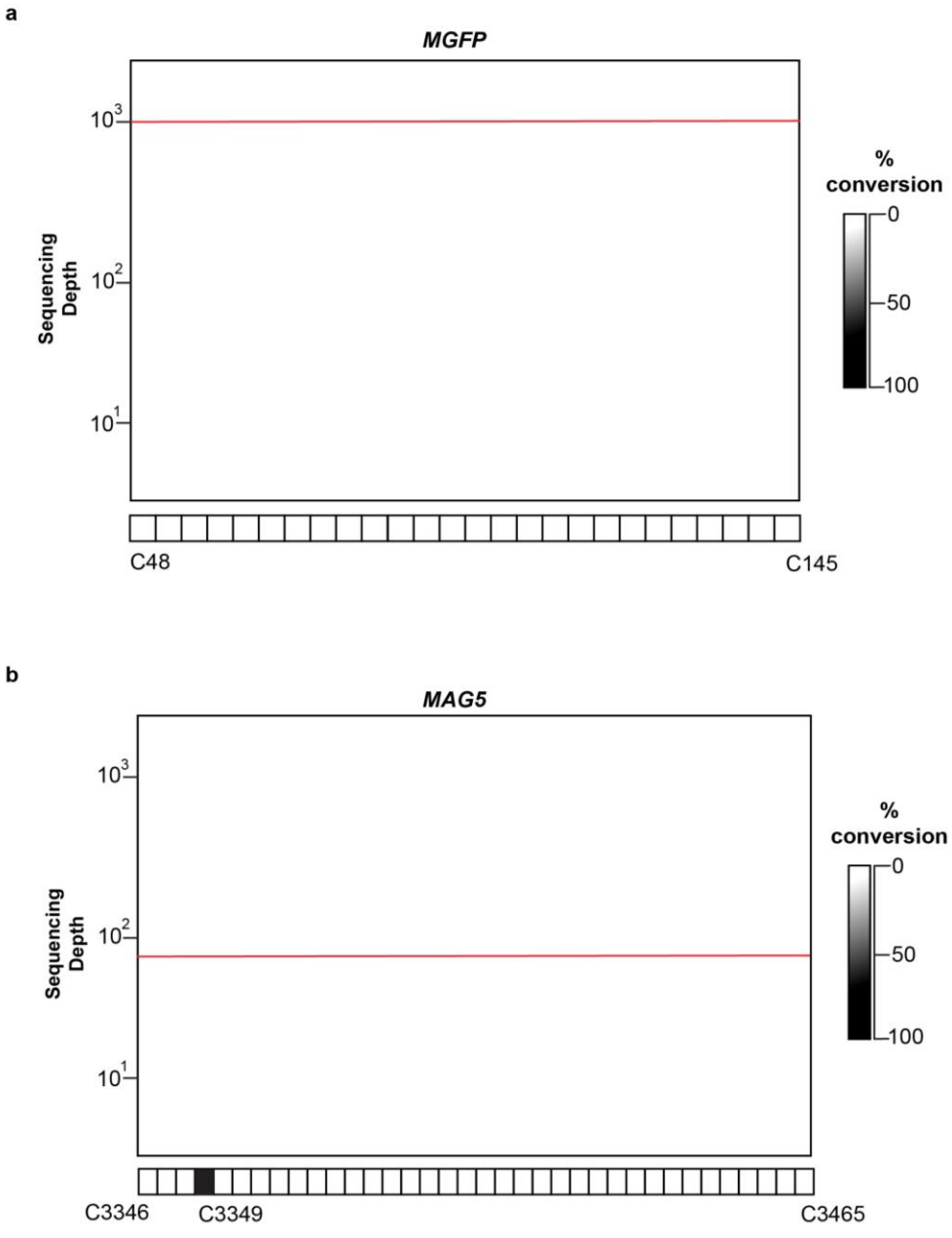


Fig 4. Representative analysis of an Illumina MiSeq amplicon sequencing of a negative and positive controls. (**a**) A region of the MGFP spiked-in *in vitro* control transcript showing even coverage and all cytosines converted (no methylation). The y axis shows the read depth and the x axis shows the cytosines (numbers) in the sequenced region. (**b**) A region of the *Mag5* gene that shows converted and a non-verted cytosine, C3349. Cytosines flanking C339 are completely converted, demonstrating that bisulphite conversion was very efficient. The heatmaps display the cytosine non-conversion percentage.

Notes

1. Slowly add 10 M NaOH dropwise to the sodium bisulfite solution while mixing. Slightly less than 1 mL is required to adjust the pH to 5.1.
2. The MiSeq Reagent Kits v3 (150 or 600-cycle) provides either 1 x 150 bp or the 600-cycle kit allows combinations of cycles that add to 600, for example 200 and 400 cycles.
3. Do not machine or finger vortex the RNA as this will increase the risk of RNA loss.
4. Air-drying the samples in a sterile laminar flow hood is best. Do not allow the RNA to completely dry as this will cause difficulties in re-suspending the RNA.
5. As the *in vitro* MGFP transcript will most likely be at a high concentration, it is good practice to perform a serial dilution in H_2O such that the estimated concentrations are in the range of 5-50 ng/μL. Prepare and run 3 dilutions on the RNA Nano chip.
6. Tilt the 1.5 mL tube at a 45^o angle and then slowly pipette the mineral oil directly on top of the RNA-bisulphite reaction mixture.

7. Emptying of the Micro Bio-Spin gel column takes about 2 min. If the gel column does not empty by gravity, place the lid back onto the column and remove again.
8. Gently pipette the reaction mixture onto the gel bed and avoid disturbing the gel bed. Minimize the transfer of mineral oil to the column although there will be traces which is unavoidable.
9. About 25% of the RNA is lost during the procedure, and we find that 10 µL of H_2O/2 µg RNA used in the bisulfite conversion reaction results in concentrations of ~150 ng/µL.
10. Bisulphite treatment of the RNA causes significant shearing and we have observed that shorter amplicons are preferentially amplified over longer amplicons.
11. Inefficient bisulfite conversion may result in unconverted cytosines, so it is important to ensure the PCR primers are not biasing the amplification towards converted cytosines.
12. Longer PCR amplicons increase the tendency of detecting non-converted cytosines in RNA exhibiting strong secondary structure.
13. Occasionally, not all triplicates successfully amplify and it may be necessary to optimise the PCR.
14. We elute the purified PCR products in 10-30 µL depending on the amount of amplified PCR products.
15. After purification of the amplicons, residual ethanol may remain in the purified amplicons. We find that concentrating down the pooled amplicons even if there is <55 µL and addition of H2O to 55 µL is best to remove as much ethanol as possible.

16. The sample sheet is required to insert the sample names and adaptor indices used for each sample. We have selected the “Other” as the category followed by “Fastq only”. This option generates fastq files only and also enables the deselection of down-stream processing steps like adaptor trimming, allowing trimming and mapping to be performed separately.
17. The prepared PhiX library is added to the pooled amplicon libraries as an internal control for the MiSeq sequencing run.
18. It is best to prepare fresh 0.2 M NaOH for the denaturation of libraries.
19. Loading 10 % PhiX control library is sufficient for low-diversity libraries. We have previously loaded between 7 to 10 pM. Under loading of the libraries can give cluster densities below the optimal range and overloading of the libraries can give cluster densities above the optimal range, reducing the quality of the data. The optimal cluster density is 700–1000 K/mm^2.

Acknowledgements

This work was supported by an Australian Research Council Future Fellowship (FT130100525) awarded to IS, a Grains Research and Development Corporation scholarship awarded to AB, a Chinese Scholarship Council scholarship awarded to JZ and a MOET-VIED PhD scholarship awarded to TD.

References

1. Burgess A, David R, Searle IR (2016) Deciphering the epitranscriptome: A green perspective. J Integr Plant Biol 58 (10):822-835.

2. Saletore Y, Meyer K, Korlach J, Vilfan ID, Jaffrey S, Mason CE (2012) The birth of the Epitranscriptome: deciphering the function of RNA modifications. Genome Biol 13 (10):175.
3. Bujnicki JM (2001) In silico analysis of the tRNA: m1A58 methyltransferase family: homology-based fold prediction and identification of new members from Eubacteria and Archaea. FEBS Lett 507 (2):123-127.
4. Droogmans L, Roovers M, Bujnicki JM, Tricot C, Hartsch T, Stalon V, Grosjean H (2003) Cloning and characterization of tRNA (m1A58) methyltransferase (TrmI) from Thermus thermophilus HB27, a protein required for cell growth at extreme temperatures. Nucleic Acids Res 31 (8):2148-2156.
5. Oerum S, Dégut C, Barraud P, Tisné C (2017) m1A Post-Transcriptional Modification in tRNAs. Biomolecules 7 (1):20.
6. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature 485 (7397):201-206.
7. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell 149 (7):1635-1646.
8. Squires JE, Patel HR, Nousch M, Sibbritt T, Humphreys DT, Parker BJ, Suter CM, Preiss T (2012) Widespread occurrence of 5-methylcytosine in human coding and noncoding RNA. Nucleic Acids Res:gks144.

9. Lovejoy AF, Riordan DP, Brown PO (2014) Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. PloS one 9 (10):e110799.
10. Khoddami V, Cairns BR (2013) Identification of direct targets and modified bases of RNA cytosine methyltransferases. Nat Biotechnol 31 (5):458-464.
11. Hussain S, Sajini AA, Blanco S, Dietmann S, Lombard P, Sugimoto Y, Paramor M, Gleeson JG, Odom DT, Ule J (2013) NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. Cell reports 4 (2):255-261.
12. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature 452 (7184):215-219.
13. Schaefer M, Pollex T, Hanna K, Lyko F (2008) RNA cytosine methylation analysis by bisulfite sequencing. Nucleic Acids Res 37 (2):e12-e12.
14. David R, Burgess A, Parker B, Li J, Pulsford K, Sibbritt T, Preiss T, Searle IR (2017) Transcriptome-Wide Mapping of RNA 5-Methylcytosine in *Arabidopsis* mRNAs and Noncoding RNAs. The Plant Cell 29 (3):445-460.
15. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (15):2114-2120.
16. Rieder D, Amort T, Kugler E, Lusser A, Trajanoski Z (2015) meRanTK: methylated RNA analysis ToolKit. Bioinformatics 32 (5):782-785.

Chapter 4: Transposable Elements (TEs) Contribute to Stress-related Long Intergenic Noncoding RNAs in Plants

Dong Wang[1], Zhipeng Qu[2], Lan Yang[1], Qingzhu Zhang[1], Zhi-Hong Liu[1], Trung Do[2], David L. Adelson[2], Zhen-Yu Wang[3], Iain Searle[2,*], and Jian-Kang Zhu[1,4]

[1]Shanghai Center for Plant Stress Biology, Shanghai Institute for Biological Science, Chinese Academy of Sciences, Shanghai 200032, China,

[2]Department of Molecular and Biomedical Sciences, School of Biological Sciences, The University of Adelaide, Adelaide, South Australia, 5005, Australia,

[3]Hainan Key laboratory for Sustainable Utilization of Tropical Bioresources, College of Agriculture, Hainan University, Haikou, China, and

[4]Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN 47907, USA

[*] Corresponding author: iain.searle@adelaide.edu.au

Submitted to Plant Journal, 27 October 2016; revised 1 January 2017; accepted 5 January 2017; published online 20 January 2017.

Statement of Authorship

Title of paper	Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in plants
Publication Status	Published
Publication details	Wang D, Qu Z, Yang L, Zhang Q, Liu ZH, Do T, Adelson DL, Wang ZY, Searle I, Zhu JK, 2017, "Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in Plants", *Plant J.* 90: 133-146.

Author Contributions

By signing the Statement of Authorship, each author certifies that:

vii. The candidate's stated contribution to the publication is accurate (as detailed below)
viii. Permission is granted for the candidate in include the publication in the thesis and
ix. The sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Candidate	Do, T		
Contribution to paper	Edited the manuscript and composed the Figure 5a.		
Overall percentage (%)	5%		
Signature		Date	15/03/2018

Name of Co-Author	Wang, D		
Contribution to paper	Performed experiments. Wrote and edited manuscript. Overall contribution 35%.		
Signature		Date	16/03/2018

Name of Co-Author	Qu, Z		
Contribution to paper	Performed experiments. Wrote and edited manuscript. Overall contribution 35%.		
Signature		Date	15/03/2018

Name of Co-Author	Yang, L		
Contribution to paper	Performed some experiments and edited manuscript. Overall contribution 5%.		
Signature		Date	11/09/18

Name of Co-Author	Zhang, Q		
Contribution to paper	Performed some experiments and edited manuscript. Overall contribution 5%.		
Signature		Date	08/09/18

Name of Co-Author	Liu, ZH		
Contribution to paper	Performed some experiments and edited manuscript. Overall contribution 5%.		
Signature		Date	16/03/2018

Name of Co-Author	Adelson, DL		
Contribution to paper	Interpreted results and edited manuscript. Overall contribution 2.5%.		
Signature		Date	15/3/2018

Name of Co-Author	Wang, ZY		
Contribution to paper	Interpreted results and edited manuscript. Overall contribution 2.5%.		
Signature		Date	08/04/18

Name of Co-Author	Searle, I		
Contribution to paper	Conceived project, designed experiments and edited manuscript. Overall contribution 2.5%.		
Signature		Date	15/3/18

Name of Co-Author	Zhu, JK		
Contribution to paper	Interpreted results and edited manuscript. Overall contribution 2.5%.		
Signature		Date	06/09/2018

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2017) **90**, 133–146 doi: 10.1111/tpj.13481

Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in plants

Dong Wang[1,†], Zhipeng Qu[2,†], Lan Yang[1], Qingzhu Zhang[1], Zhi-Hong Liu[1], Trung Do[2], David L. Adelson[2], Zhen-Yu Wang[3], Iain Searle[2,*] and Jian-Kang Zhu[1,4,*]

[1]*Shanghai Center for Plant Stress Biology, Shanghai Institute for Biological Science, Chinese Academy of Sciences, Shanghai 200032, China,*

[2]*Department of Genetics and Evolution, School of Biological Sciences, The University of Adelaide, Adelaide, South Australia, 5005, Australia,*

[3]*Hainan Key laboratory for Sustainable Utilization of Tropical Bioresources, College of Agriculture, Hainan University, Haikou, China, and*

[4]*Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN 47907, USA*

Received 27 October 2016; revised 1 January 2017; accepted 5 January 2017; published online 20 January 2017.
*For correspondence (e-mails jkzhu@sibs.ac.cn or iain.searle@adelaide.edu.au).
†These authors contributed equally to this work.

SUMMARY

Noncoding RNAs have been extensively described in plant and animal transcriptomes by using high-throughput sequencing technology. Of these noncoding RNAs, a growing number of long intergenic noncoding RNAs (lincRNAs) have been described in multicellular organisms, however the origins and functions of many lincRNAs remain to be explored. In many eukaryotic genomes, transposable elements (TEs) are widely distributed and often account for large fractions of plant and animal genomes yet the contribution of TEs to lincRNAs is largely unknown. By using strand-specific RNA-sequencing, we profiled the expression patterns of lincRNAs in Arabidopsis, rice and maize, and identified 47 611 and 398 TE-associated lincRNAs (TE-lincRNAs), respectively. TE-lincRNAs were more often derived from retrotransposons than DNA transposons and as retrotransposon copy number in both rice and maize genomes so did TE-lincRNAs. We validated the expression of these TE-lincRNAs by strand-specific RT-PCR and also demonstrated tissue-specific transcription and stress-induced TE-lincRNAs either after salt, abscisic acid (ABA) or cold treatments. For Arabidopsis TE-lincRNA11195, mutants had reduced sensitivity to ABA as demonstrated by longer roots and higher shoot biomass when compared to wild-type. Finally, by altering the chromatin state in the Arabidopsis chromatin remodelling mutant *ddm1*, unique lincRNAs including TE-lincRNAs were generated from the preceding untranscribed regions and interestingly inherited in a wild-type background in subsequent generations. Our findings not only demonstrate that TE-associated lincRNAs play important roles in plant abiotic stress responses but lincRNAs and TE-lincRNAs might act as an adaptive reservoir in eukaryotes.

Keywords: transposable element, long intergenic noncoding RNAs, transposable elements-associated lincRNAs, abiotic stress, noncoding RNAs.

INTRODUCTION

Noncoding RNAs (ncRNA) have been extensively described in plant and animal transcriptomes by using high-throughput sequencing technology. Besides canonical ncRNAs that include ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small nuclear and small nucleolar RNAs, many regulatory ncRNAs have been characterized (Cech and Steitz, 2014). Small regulatory RNAs, including microRNAs and small interfering RNAs (siRNAs), have been demonstrated to play important roles in the regulation of eukaryotic gene expression through either transcriptional or post-transcriptional mechanisms (Bologna and Voinnet, 2014). These small RNAs are produced by cleavage of folded double-stranded RNA (dsRNA) derived from long noncoding RNA (lncRNA). A growing number of lncRNAs have been shown

to function in gene regulation without being processed into small RNAs. In animals, to balance the copy number of X chromosomes between male and female cells, the lncRNA Xist recruits Polycomb group proteins to cause lysine 27 trimethylation in histone H3 (H3K27me3) to silence one X chromosome in females (Plath *et al.*, 2003). In plants, thousands of lncRNAs generated by the DNA-dependent RNA polymerase V are involved in RNA-directed DNA methylation and transcriptional gene silencing (Wierzbicki *et al.*, 2008).

Recently, one type of lncRNA, long intergenic noncoding RNAs (lincRNAs), have been identified by tiling arrays or RNA-sequencing in several plant species (Liu *et al.*, 2012; Li *et al.*, 2014; Zhang *et al.*, 2014). LincRNAs are defined as ncRNA longer than 200 nt that do not overlap with either protein-coding or other non-lincRNA types of genes (Ulitsky and Bartel, 2013). Some of them are known to play fundamental biological roles in plant development and physiology (Ariel *et al.*, 2014, 2015; Zhang *et al.*, 2014), such as *INDUCED BY PHOSHATE STARVATION1* (*IPS1*), that can inhibit the function of miR319 through target mimicry during inorganic phosphate starvation response (Franco-Zorrilla *et al.*, 2007). Although the functions of lincRNAs are beginning to be studied in plants, their origin still remains obscure.

Transposable elements (TEs) have been found to be widely distributed in many eukaryotic genomes, and constitute a large fraction of plant and animal genomes. In humans, more than two-thirds of mature lncRNAs contain an exon of at least partial TE origin (Kapusta *et al.*, 2013), and they are believed to contribute contemporary sequence elements to conserved lncRNAs in animals (Hezroni *et al.*, 2015). The contribution of TEs to lincRNA in plants is still unknown. In this report, we explored the contribution of TEs to lincRNAs in three plants species, with significantly different genomic TE diversity. Proportions of lincRNAs harbouring TEs are significantly higher in maize and rice than in Arabidopsis, which is consistent with the number of TEs in these genomes. We name these lincRNAs containing TEs, TE-associated lincRNAs (TE-lincRNAs), and show that some of them are expressed in a tissue-specific pattern. Of particular interest was the observation that the expression pattern of some TE-lincRNAs varied in response to different stress conditions. Furthermore, *Arabidopsis thaliana* seedlings deficient in TE-lincRNA11195, were more resistant to abscisic acid (ABA) treatment when compared to wild-type (WT), indicating that this lincRNA was involved in the abiotic stress response. Importantly, unique lincRNAs, including TE-lincRNAs, were transcribed in seedlings with DDM1 (decrease in DNA methylation 1) loss of function, and these lincRNAs were inherited in subsequent generations in a WT background, suggesting that these unique lincRNAs produced by changing the chromatin status can be inherited.

RESULTS

Genome-wide identification of TE-lincRNAs in three plant species

To systematically identify TE-lincRNAs, we performed strand-specific RNA-sequencing from 2-week-old seedlings of three plant species. Because of the low expression levels of retrotransposon-derived lncRNAs reported in human and mouse (Fort *et al.*, 2014), we produced high-depth transcriptomes, of approximately 66 million, 173 million, and 256 million pair-end Illumina reads from three biological replicates of *Arabidopsis thaliana*, rice (*Oryza sativa* subsp. *japonica*) and maize (*Zea mays* subsp. *mays* var. B73), respectively (Table S1). We constructed a comprehensive pipeline to identify TE-lincRNAs, consisting of three key steps (Figure 1). First, transcripts from the three species were reconstructed from their RNA-seq datasets using Cufflinks (Trapnell *et al.*, 2010) after mapping reads to the corresponding reference genomes with TopHat2 (Kim *et al.*, 2013). Second, only transcripts greater than 200 nt and not overlapping annotated genes were kept, and we then removed potentially peptide/protein-coding transcripts by sequence similarity search against SWISSPROT and filtered out transcripts with open reading frames (ORFs) larger than 100 amino acids (aa) inside or 50 aa at end(s). After filtering, 205, 1229 and 773 transcripts remained, corresponding to lincRNAs in Arabidopsis, rice and maize, respectively (Table 1). Third, lincRNAs partially overlapping TE loci but not completely located inside TEs were classified as TE-lincRNAs. In the end, we identified 47, 611 and 398 TE-lincRNAs from Arabidopsis, rice and maize, respectively (Tables 1 and S2). The significantly larger proportion of TE-lincRNAs in rice and maize when compared to Arabidopsis is correlated with the increased number of TEs (Table 1). We then determined the genomic distribution of TE-lincRNAs in all three genomes and found that TE-lincRNAs were distributed on all nuclear chromosomes, but were not strongly correlated with the distributions of TEs along the chromosome (Figure S1).

We then compared some general characteristics of TE-lincRNAs and lincRNAs without TEs (designated as non-TE-lincRNAs). Both TE-lincRNAs and non-TE-lincRNAs share similar length distributions in all three species (Figure S2), while the average lengths of TE-lincRNAs are significantly longer than non-TE-lincRNAs in Arabidopsis (829 nt compared to 773 nt, P-value = 0.01347, Wilcoxon rank sum test), rice (1125 nt compared to 834 nt, P-value < 2.2e-16, Wilcoxon rank sum test) and maize (1343 nt compared to 753 nt, P-value < 2.2e-16, Wilcoxon rank sum test). The majority of lincRNAs, including TE-lincRNAs and non-TE-lincRNAs, are single-exon transcripts in all three species examined (Figure S3). There is no significant difference between the average exon numbers of

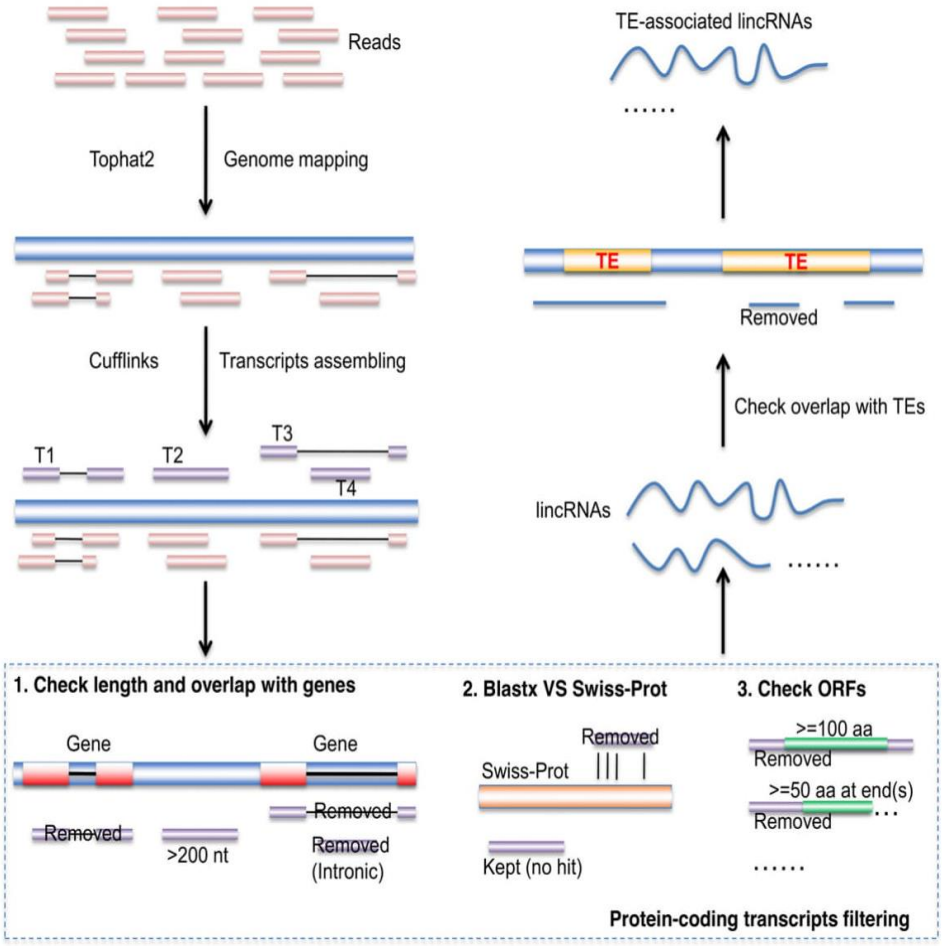


Figure 1. Identification of TE-associated lincRNAs from RNA-seq data.
Quality-checked short reads were mapped to the reference genome using TopHat2 and Cufflinks was then used to assemble the mapped reads into longer transcripts. To filter out protein-coding transcripts and canonical noncoding RNA the following three steps were undertaken. First, transcripts shorter than 200 nt were removed and the remaining were tested for overlap with annotated genes. Those transcripts that either overlapped with annotated genes by at least one base pair or that were located in the intronic regions of genes were removed. Second, transcripts with high similarity to known protein motifs were identified by BLASTX searches against the SWISS_PROT database and then removed. The last step involved inspecting the transcript ORFs and removing transcripts with ORFs longer than 100 amino acids (aa) inside the transcript or longer that 50 aa at transcript end(s). These remaining transcripts were classified as candidate lincRNAs. TE-associated lincRNAs were identified as those that overlapped with transposable element (TE) loci but did not fully reside within a TE. [Colour figure can be viewed at wileyonlinelibrary.com].

Table 1 Summary of lincRNAs identified in this study

Species	Number of total lincRNAs	Number of TE-associated lincRNAs	Proportion of transposable elements in genome (%)	Proportion of TE-associated lincRNAs in total lincRNAs (%)
A. thaliana	205	47	14	22.9
O. sativa subsp. *japonica*	1229	611	35	49.7
Z. mays B73	773	398	76	51.5
A. thaliana (*ddm1* mutant)	446	102	14	22.9

TE-lincRNAs and non-TE-lincRNAs in Arabidopsis and rice, while only slightly significant lower average exon numbers for TE-lincRNAs in maize (1.6 compared to 1.5, *P*-value = 0.2507 in Arabidopsis; 1.6 compared to 1.7, *P*-value = 0.1432 in rice; 1.3 compared to 1.4, *P*-value = 0.007197 in maize; Wilcoxon rank sum test). These results indicated that TEs may have contributed to the extension of transcribed length of lincRNAs but not to splicing complexity in rice and maize. In addition, we scored the potential of RNA motifs embedded in TE-lincRNAs and non-TE-

lincRNAs by utilizing the Rfam database, and most lincRNAs, either TE-lincRNAs or non-TE-lincRNAs, have none or only one RNA motif (Figure S4 and Table S3). There was no significant difference with respect to the number of embedded RNA motifs between TE-lincRNAs and non-TE-lincRNAs (*P*-value = 0.8368 in Arabidopsis; *P*-value = 0.5387 in rice; *P*-value = 0.8285 in maize; Wilcoxon rank sum test). Next we determined if positional bias of lincRNAs with respect to corresponding neighboring protein-coding genes occurs in the three genomes. Both TE-lincRNAs and non-TE-lincRNAs showed biased distributions at 5′ or 3′ end 5 kilobase (kb) flanking regions of protein-coding genes (Figure S5). We also checked the correlation of expression profiles of TE-lincRNAs and non-TE-lincRNAs with their 10 closest genes at the 5′ end or 3′ end using public RNA-seq datasets (Figure S6a) (Filichkin *et al.*, 2010; Di *et al.*, 2014). We observed the significant high positive or negative expression correlation between some TE-lincRNAs or non-TE-lncRNAs with their neighbor genes, but not for all lincRNAs. Then we reconstructed the protein-coding and non-protein-coding RNA co-expression networks based on the expression profiles across these RNA-seq datasets, and 16 320 genes as well as 77 lincRNAs (including 12 TE-lincRNAs) were reconstructed into 21 co-expression sub-networks (Table S4). TE-lincRNAs were identified with high expression correlation with multiple protein-coding genes in co-expression sub-networks showing stress response (Figure S6b, c).

Examination of TE contributions to lincRNAs

Plant TEs are primarily of two types: class I (retroelement) transposing through an RNA intermediate (copy and paste mechanism) and class II (DNA element) using a DNA intermediate (cut and paste mechanism) to transpose (Bennetzen and Wang, 2014). These two types of TEs can be further classified into many families based on their sequence similarity (Wicker *et al.*, 2007), and each family of TEs has its own functional properties and evolutionary history. Therefore, we were interested in studying the contribution of different TE families to lincRNAs. In Arabidopsis, more than 40% of TE-lincRNAs (22 out of 47) contained 28 RC/Helitron TEs (Figure 2a and Table S5). In rice, the

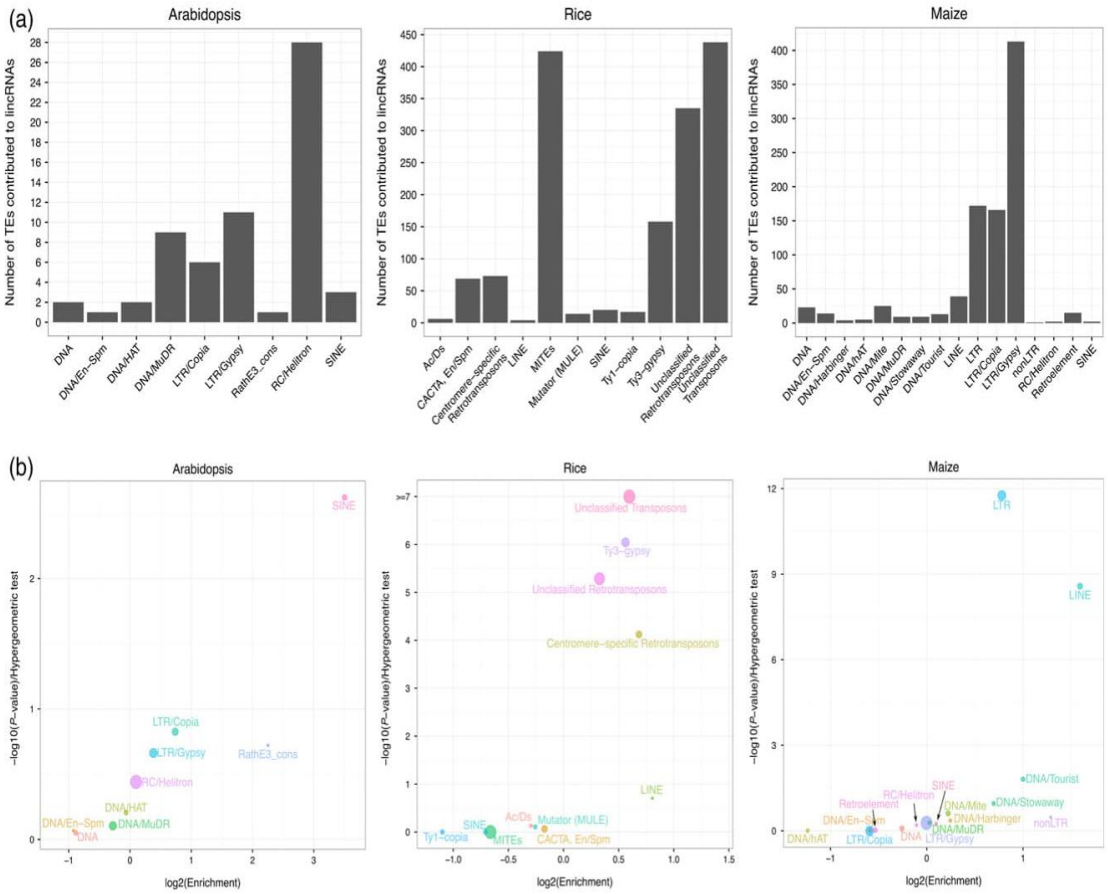


Figure 2. Occurrence and enrichment of different TE families in lincRNAs from Arabidopsis, rice and maize.
(a) Bar charts showing the number of TEs from different families contributing to lincRNAs.
(b) Bubble charts describing the over-representation of different TE families contributing to TE-associated lincRNAs. X axis represents the fold of enrichment of different TE families contributing to lincRNAs. Y axis represents statistical significance of the over-representation of different TE families contributing to lincRNAs (*P*-value, hypergeometric test). Sizes of bubbles indicate proportions of TEs in each TE family with respect to total number of TEs contributing to lincRNAs. [Colour figure can be viewed at wileyonlinelibrary.com].

majority of TE-lincRNAs (228, 247 and 197 out of 611) harboured 424 miniature inverted-repeat transposable elements (MITEs), 438 unclassified transposons and 335 unclassified retrotransposons, respectively. While in maize, Gypsy (413 Gypsys contributed to 230 lincRNAs), LTR (172 LTRs contributed to 116 lincRNAs) and Copia (166 Copias contributed to 113 lincRNAs) retrotransposons were the three major contributors in inbred line B73 (Figure 2a and Table S5). Because the copy number of different TEs in these genomes differs, enrichment analysis was carried out to examine the significance of contributions of different TEs to lincRNAs. The short interspersed elements (SINEs) in Arabidopsis made the most significant contribution to lincRNAs (Figure 2b). While unclassified transposons and unclassified retrotransposons, Ty3-gypsy and centromere-specific retrotransposons contributed remarkably to rice lincRNAs (Figure 2b). In maize, LTR and long interspersed elements (LINE) were most significant enriched TE families in lincRNAs (Figure 2b). Aside from TE families over-represented in their contribution to lincRNAs, there were some TE families under-represented in their contribution to lincRNAs. Nine TE families of Arabidopsis were excluded from lincRNA transcripts, including LINE/L1 with a copy number greater than 1000 (Table S5). In rice, no lincRNAs harboured segments of Mariner while DNA/hAT-Ac with a copy number of approximately 3200 was one of five TE families that did not contribute lincRNAs in maize (Table S5). These results suggest that different TE families have different contributions to lincRNAs in varied plant species. Compared to the two crops used in this study, more TE families tended to be depleted from lincRNAs in 2-week-old Arabidopsis seedlings.

With respect to the number of TEs contributing to individual TE-lincRNA, we found that the largest number of lincRNAs contain only a single TE, while some of TE-lincRNAs can contain up to 18 or 43 TEs, in rice and maize respectively (Figure S7(a)). Length of TE-lincRNAs contributed by more than one TEs are longer than those contributed by one TE (Figure S7b). When considering the coverage of lincRNAs contributed to by TEs, we found that many lincRNAs had a high percentage of TE content, especially in maize (Figure S7c). Conversely, most TEs that contributed to lincRNAs are fully inside TE-lincRNAs (Figure S7d). We also found that the percentage of TE-lincRNAs in identified lincRNAs was much higher than the percentage of genes contributed to by TEs (Figure S8a). Specifically, TE coverage in TE-lincRNAs was significantly higher than TE coverage in protein-coding genes in maize (mean coverage as 54.3% to 16.7%, P-value < $2.2e^{-16}$, Wilcoxon rank sum test), but not in Arabidopsis (mean coverage as 33.4–36.9%, P-value = 0.2894, Wilcoxon rank sum test) and rice (mean coverage as 35.6–38.5%, P-value = 0.1355, Wilcoxon rank sum test) (Figure S8b). In addition, we also checked the coordinates of TEs with respect to host lincRNAs. Most TEs were completely nested inside the lincRNAs, as we have shown in Figure S7 (d), while most of the remaining TEs were located within 500-bp flanking regions of lincRNAs (Figure S7e).

We also analysed the conservation of TE-lincRNA between Arabidopsis and rice according to the protocol described in the methods, but because the number of whole-genome pairwise alignments between maize and other species was small (only four), this conservation analysis was not performed for maize. The overall conservation levels of different genomic features were similar in both Arabidopsis and rice as measured by the phyloP score. As expected, the most conserved element was genes, the least conserved was TEs and TE-lincRNAs were more conserved than TEs (Figure 3). TE-lincRNAs and non-TE-lincRNAs had a similar level of conservation (Figure 3). This was broadly consistent with the idea that TEs embedded in lncRNAs were functionally or structurally constrained by evolution (Kapusta *et al.*, 2013).

Transcript profiling of TE-lincRNAs in Arabidopsis

Next we validated expression of TE-lincRNAs in seedlings of Arabidopsis, maize and rice by strand-specific reverse transcription (RT)-PCR. We selected 11 candidates for further expression analysis in Arabidopsis of which we validated expression for all of them (Figure 4a). All TE-lincRNAs tested were amplified from only one strand as expected from our directional RNA-seq data. Moreover, they were amplified from cDNA primed with oligodT indicating that the TE-lincRNAs were polyadenylated. Similarly in rice and maize, all three TE-lincRNAs from each species were confirmed to be expressed as all were amplified from strand-specific cDNA or oligodT primed cDNA (Figure 4b, c). To measure TE-lincRNAs transcript levels in different Arabidopsis tissues we performed digital PCR on a Fluidigm Biomark HD system. All TE-lincRNAs exhibited varied expression patterns in different tissues. For example, lincRNA18980 was found to be highly expressed in roots but almost not expressed in flowers, and TE-lincRNA3688, TE-lincRNA11344 and TE-lincRNA15772 showed very low levels of expression in root, flower and silique tissues (Figure 5(a)). In addition, transcript profiles of 205 Arabidopsis lincRNAs under different stress treatments were analysed using public RNA-seq data (Filichkin *et al.*, 2010). Compared with normal growth conditions, the expression patterns of many lincRNAs, including TE-lincRNAs, were altered in five stress conditions (Figure 5b). This observation was consistent with early studies that lncRNAs exhibit tissue-specific or spatiotemporal patterns (Cabili *et al.*, 2011; Derrien *et al.*, 2012; Goff *et al.*, 2015). Because many lncRNAs have been shown to be involved in gene regulation *in cis*, we further checked the correlation of expression between selected TE-lincRNAs and their neighbouring genes. For TE-lincRNA15772 and TE-lincRNA19433, there

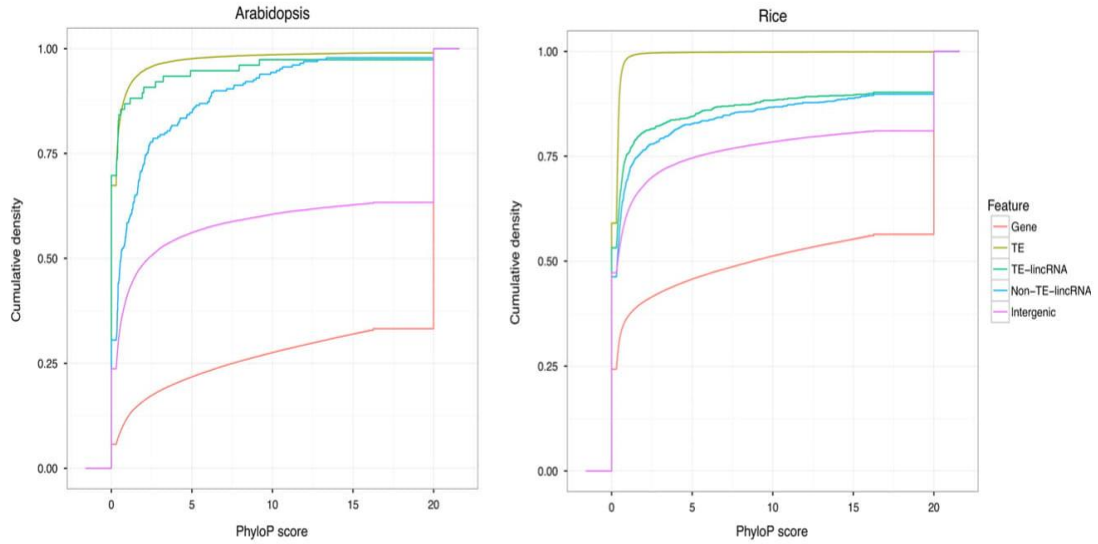


Figure 3. Level of conservation of TE-lincRNAs, non-TE-lincRNAs, genes, TEs and intergenic regions in Arabidopsis and rice.
The cumulative distributions of phyloP scores derived from 24-way (Arabidopsis) and 28-way (rice) whole-genome alignments are presented. [Colour figure can be viewed at wileyonlinelibrary.com].

was no correlation between the expression of these TE-lincRNAs and their flanking genes; however, the transcript level of TE-lincRNA11344 was negatively correlated with expression of its neighbouring gene DPBF2 (Spearman's correlation, $r = -0.9036145$, P-value = 0.00208, Figure 5c), suggesting that TE-lincRNA11344 may function by down-regulating the adjacent gene.

Mutations in Arabidopsis TE-lincRNA11195 cause resistance to abscisic acid

To investigate functional roles of TE-lincRNAs during stress conditions, we identified homozygous T-DNA insertion mutants in a number of TE-lincRNAs and screened the mutants under standard laboratory conditions and during ABA treatment (Alonso *et al.*, 2003). Strikingly, two independent T-DNA insertion alleles of TE-lincRNA11195 containing a LTR exhibited ABA resistant phenotypes (Figure 6). T-DNA insertions in both mutants caused TE-lincRNA11195 transcript to be undetectable (Figure 6a), and we designated these two lines as *11195-1* and *-2*. In the absence of exogenous ABA, *11195-1* and *11195-2* seedlings had similar growth when compared to WT (Figure 6b). However, after moving to media supplemented with 20 μM ABA, remarkably enhanced resistance was observed in the mutants compared with WT (Figure 6(b)). Both mutants had significantly increased primary root elongation when compared to WT under 20 μM ABA treatment (Figure 6(c) top panel, Two-sample independent *t*-test, $P < 0.05$), and a weak but non-significant enhancement in the fresh weight of aerial tissues (Figure 6c bottom panel, Two-sample independent *t*-test, $P > 0.05$). In addition, we tested whether TE-lincRNA11195 plays a role in seed germination. Mutants of *lincRNA11195* also showed insensitive to exogenous ABA at the stage of seed germination (Figure S9a), and were substantially insensitive to ABA in post-germination seedling development (Figure S9b, c, Two-sample independent *t*-test, $P < 0.01$).

To investigate the transcription regulation of TE-lincRNA11195, we measured TE-lincRNA11195 RNA abundance in WT and ABA insensitive mutants by RT-PCR. Abundance of lincRNA11195 increased more than two-fold under ABA treatment at 12 h in Col-0 (Figure S10). Mutant seedlings of the ABA receptors PYR1/PYL1/PYL4 inhibited the transcription of TE-lincRNA11195 (Figure S10). Together these findings clearly demonstrate that TE-lincRNA11195 is ABA responsive. To further explore the regulation and functional role of TE-lincRNA11195 during abiotic stress responses, TE-lincRNA11195 abundance was monitored in several stress conditions. Besides ABA treatment, salt and cold treatments changed the abundance of TE-lincRNA11195, but did not affect the adjacent gene expression (Figure S11a). Next we studied the role of TE-lincRNA11195 under salt treatment at both germination and post-germination seedling development. Seed germination and greening rates of seedlings were significantly higher in *lincRNA11195* mutants than WT (Figure S11b, c), suggesting that TE-lincRNA11195 is also involved in response to salt. Together these results indicated that TE-lincRNA11195 is involved in abiotic stress responses in plants.

In order to identify potential gene targets of TE-lincRNA11195, we performed RNA-seq on wild-type and *lincRNA11195* mutants under normal and ABA treated conditions. We used a Generalized Linear Model (GLM) to identify differential expressed genes in *lincRNA11195* amongst wild-type and ABA treatments and identified 8 and 10 genes that were significantly up- or down-regulated (Benjamini-Hochberg method adjusted P-value < 0.05), respectively (Figure 7a and Table S6). Gene Ontology (GO)

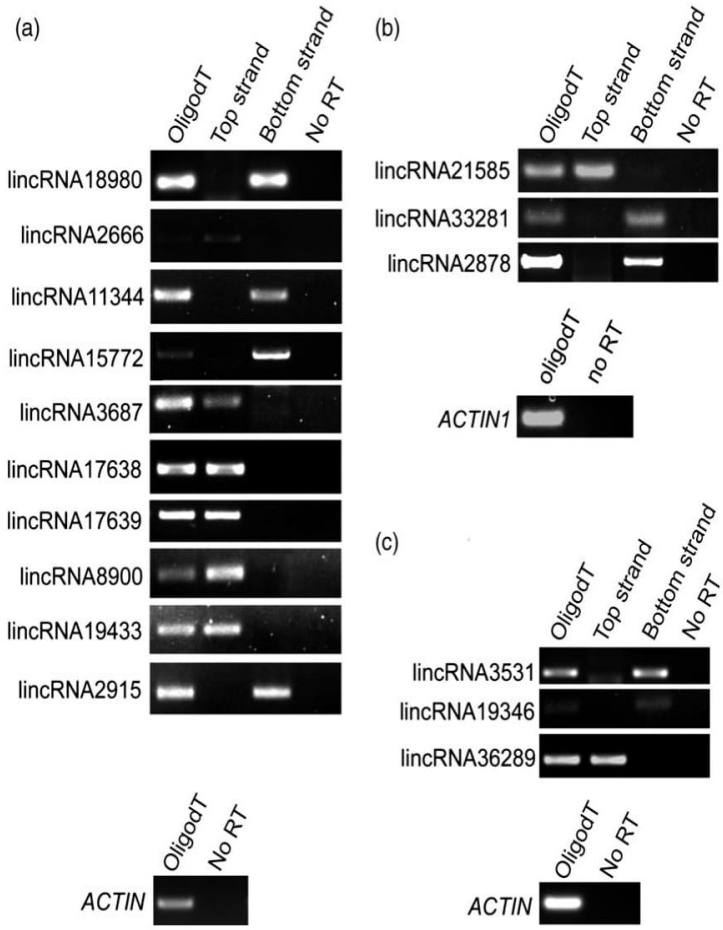


Figure 4. Detection of TE-lincRNAs in three species.
(a–c) Strand-specific RT-PCR analysis was carried out on selected TE-lincRNA transcripts in either (a) *A. thaliana*; (b) *Oryza sativa* subs. *japonica*; or (c) *Zea mays* B73. Either oligodT, top strand or bottom strand-specific primers were used in the reverse transcription cDNA synthesis. Control RT-PCRs using either *ACTIN* or *ACTIN1* primes are shown below each panel.

enrichment analysis of the 100 most significantly differentially expressed genes indicated that genes involved in 'response to salicylic acid stimulus' are most significantly over-represented (Figure 7b). The genomic distribution of these 100 most significantly differentially expressed genes showed that they are distributed across all chromosomes (Figure 7c). Further molecular analysis, for example RIP to detect RNA–protein interactions, will be required to elucidate the function of TE-lincRNA11195.

TEs insertions are known to modify transcriptional responses in plants (Naito *et al.*, 2009; Ito *et al.*, 2011), and we evaluated the contribution of the LTR to TE-lincRNA11195 transcription under ABA treatment. Expression of TE-linc11195 in transgenic plants without the LTR was slightly higher than in plants with the LTR under control conditions; but the expression of TE-linc11195 harbouring the LTR had a greater increase than in plants without the LTR under ABA treatment (Figure S12), suggesting that TE enhances the extent of TE-linc11195 ABA response at the transcriptional level. We then investigated the expression of TE-linc11195 in the close relative *Arabidopsis lyrata* and *Capsella rubella*, as the DNA sequence of TE-linc11195 is present in both species. Transcript of TE-linc11195 could be detected in two-week-old seedlings of *A. thaliana* and *A. lyrata* (Figure S13), indicating that TE-linc11195 may function specifically at this stage in the Arabidopsis genus. Next we performed a pairwise sequence alignment of TE-linc11195 between *A. thaliana* and *A. lyrata* and this indicated the majority of the sequence is conserved between these two species (Figure S14). We also performed a comparison of the secondary structures of TE-linc11195 in the two species and demonstrated they were largely conserved (Figure S14).

Characterization of unique TE-lincRNAs generated in loss of *ddm1* mutant plants

Chromatin changes can be triggered by fluctuations in the ambient environment (Talbert and Henikoff, 2014), and unique lncRNAs responsive to abiotic or biotic stress have also been characterized in plants (Di *et al.*, 2014; Zhu *et al.*, 2014). Therefore we were interested in the correlation between lincRNA expression and chromatin status in *ddm1*. Mutated chromatin-remodeling factor DDM1 alters the distribution of DNase I hypersensitive (DH) sites that are closely associated with RNA Polymerase II binding sites (Zhang *et al.*, 2012; Wang and Timmis, 2013). We generated a transcriptome dataset including approximately 70 million paired-end reads from 2-week-old *ddm1* seedlings

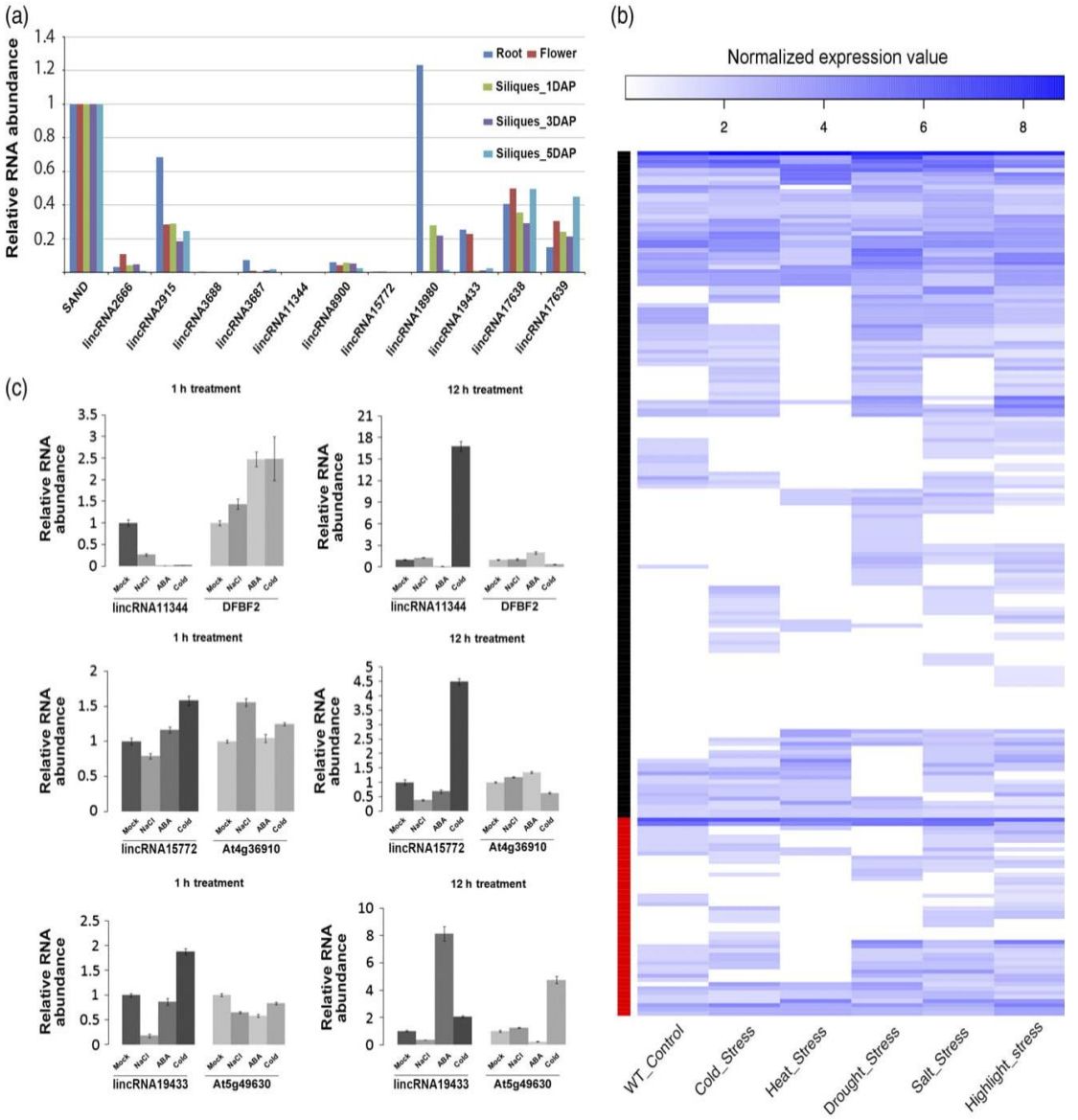


Figure 5. Expression pattern of TE-lincRNAs.

(a) Expression of TE-lincRNAs in different Arabidopsis tissues. cDNA abundance was normalized using the SAND transcript.

(b) Heatmap showing expression profiles of Arabidopsis lincRNAs under different stress conditions. Expression value was normalized by variance-stabilizing transformation of raw counts. Black sidebar: 154 non-TE-lincRNAs; red bar: 47 TE-lincRNAs.

(c) Expression of selected TE-associated lincRNAs and neighbouring genes under different conditions. *ACTIN7* was used as a control in the qRT-PCR experiments of this study. [Colour figure can be viewed at wileyonlinelibrary.com].

(Table S1). As we expected, unique transcripts were detected from intergenic regions of plants defective for DDM1, and TE-lincRNAs as well as non-TE-lincRNAs were detected (Figure 8a and Table 1). There was a similar percentage of TE-lincRNAs found in the *ddm1* lincRNA repertoire (102 out of 446) compared to WT, nonetheless the total number of TE-lincRNAs and non-TE-lincRNAs was increased in *ddm1* Col (Table 1). 387 *ddm1* specific lincRNAs were found, and 192 of them were found to be covered by DH sites by checking their position and 1-kb flanking regions (Zhang *et al.*, 2012; Wang and Timmis, 2013), indicating that unique lincRNAs can be generated once nuclear chromatin state changes. Subsequently, the inheritance of these unique lincRNAs was studied in *ddm1* heterozygous seedlings produced by crossing *ddm1* homozygous plants with WT and by intercrossing the F1 to produce F2 plants (Figure 8b). Interestingly, transcripts of these lincRNAs could be detected in heterozygous F1 seedlings (Figure 8c) and strikingly in the subsequent F2 generation expression was independent of the *DDM1* genotype (Figure 8c and Table S7). Of interest, these *ddm1* specific lincRNAs were not expressed in some of *ddm1* homozygous seedlings, indicating that the inheritance of lincRNA is non-Mendelian (Table S7).

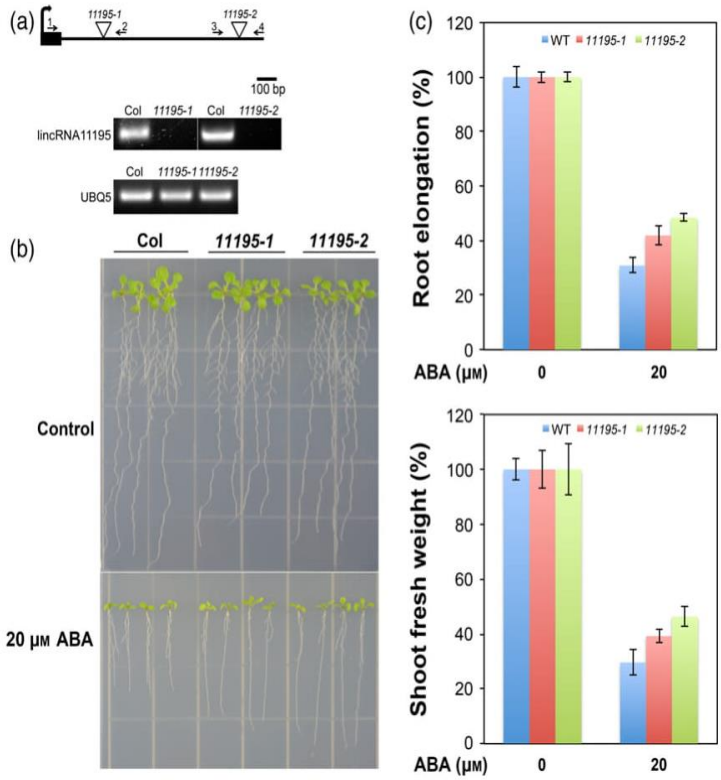


Figure 6. Arabidopsis TE-associated lincRNA11195 mediates ABA responses.

(a) Expression analysis of the TE-associated lincRNA11195 in wild-type (Col-0) and two T-DNA insertion mutant alleles (lincRNA*11195-1* and *11195-2*). Bold right curved arrow shows the direction of transcription of lincRNA11195. The two primer pairs shown (1 and 2 for *11195-1* and 3 and 4 for *11195-2*) were used to amplify the TE-lincRNA11195 transcript. UBQ5 was used as a positive control.

(b) *TE-linc11195* mutants are insensitive to ABA.

(c) Root length and fresh shoot weight of seedlings shown in (b). Both graphs are presented as the percentage relative to growth on control half-strength MS medium. ABA assays were performed by stratifying seeds at 4°C for 3 days, followed by growth of seedlings for 5 days on half MS media, then seedlings were transferred onto half-strength MS medium supplemented with or without 20 μM ABA, and grown for an additional 8 days. Error bars stand for standard deviation ($n = 20$). [Colour figure can be viewed at wileyonlinelibrary.com].

DISCUSSION

The importance of lncRNAs involved in biological processes has been extensively described in many plant species including crops thanks to advances in DNA sequencing technology (Li *et al.*, 2014; Zhang *et al.*, 2014; Wang *et al.*, 2015,2016), but comprehensive understanding of their biological function and origin still remain elusive. Although TEs are proposed to be a major contributor to vertebrate lncRNAs (Kelley and Rinn, 2012; Kapusta *et al.*, 2013), their contribution to plant lncRNAs remains unclear. In this study, we mainly focused on the contribution of TEs to lincRNA in three plant species, one model dicotyledonous species (*Arabidopsis thaliana*) and two important monocotyledonous crops (rice and maize). In total, 47, 611 and 398 TE-lincRNAs were identified in Arabidopsis, rice and maize respectively by using high-quality RNA-seq data with high-depth stranded RNA-sequencing. In rice and maize, TEs occurred in approximately half of the lincRNAs identified from 2-week-old seedlings. Despite the small proportion of TEs in the *Arabidopsis thaliana* genome, more than 20% of identified lincRNAs included TEs. This demonstrates that TEs make a remarkable contribution to lincRNAs in plants particularly as TEs constitute the majority of DNA in many plant genomes. Furthermore, lincRNAs preferentially harbour TEs compared to protein-coding genes, which is an observation that is consistent with findings in mammals (Kapusta *et al.*, 2013).

While TEs are ubiquitous in lincRNAs from all three examined plants, some TE families are excluded from the lincRNA repertoire (Table S5). Moreover, the relative abundance of TE families within lincRNAs does not simply mirror that of the entire genome. For example, the copy number of SINEs is not high, but their contribution to lincRNA is significant in *Arabidopsis thaliana* (Figure 2b). These results show that contribution to lincRNA does not mean a close correlation with the number of TEs. Also the interspecific variations we observed in the coverage and type of TEs in lincRNAs reflect the abundance and intrinsic properties of certain TEs residing in the genome, and it further suggests that TEs play a role in the divergence of lincRNAs.

LincRNAs are known to exhibit organ-specific expression patterns in Arabidopsis (Liu *et al.*, 2012), and this pattern was also observed in TE-lincRNAs (Figure 5a). Furthermore, varied TE-lincRNAs expression was observed under different stress treatments (Figure 5b), indicating their transcription is responsive to abiotic stress. This hypothesis is supported by the ABA treatment result of TE-linc11195 There is an LTR in the 5′ terminal region of TE-linc11195, and knock out of TE-linc11195 caused an ABA insensitive phenotype for *Arabidopsis thaliana* seedlings comparing with WT (Figure 6 and Figure S9). Moreover, expression of this TE-lincRNA was completely blocked in seedlings mutated ABA receptors genes PYR1/PYL1/PYL4 (Figure S10). In addition, a salt insensitive phenotype was also

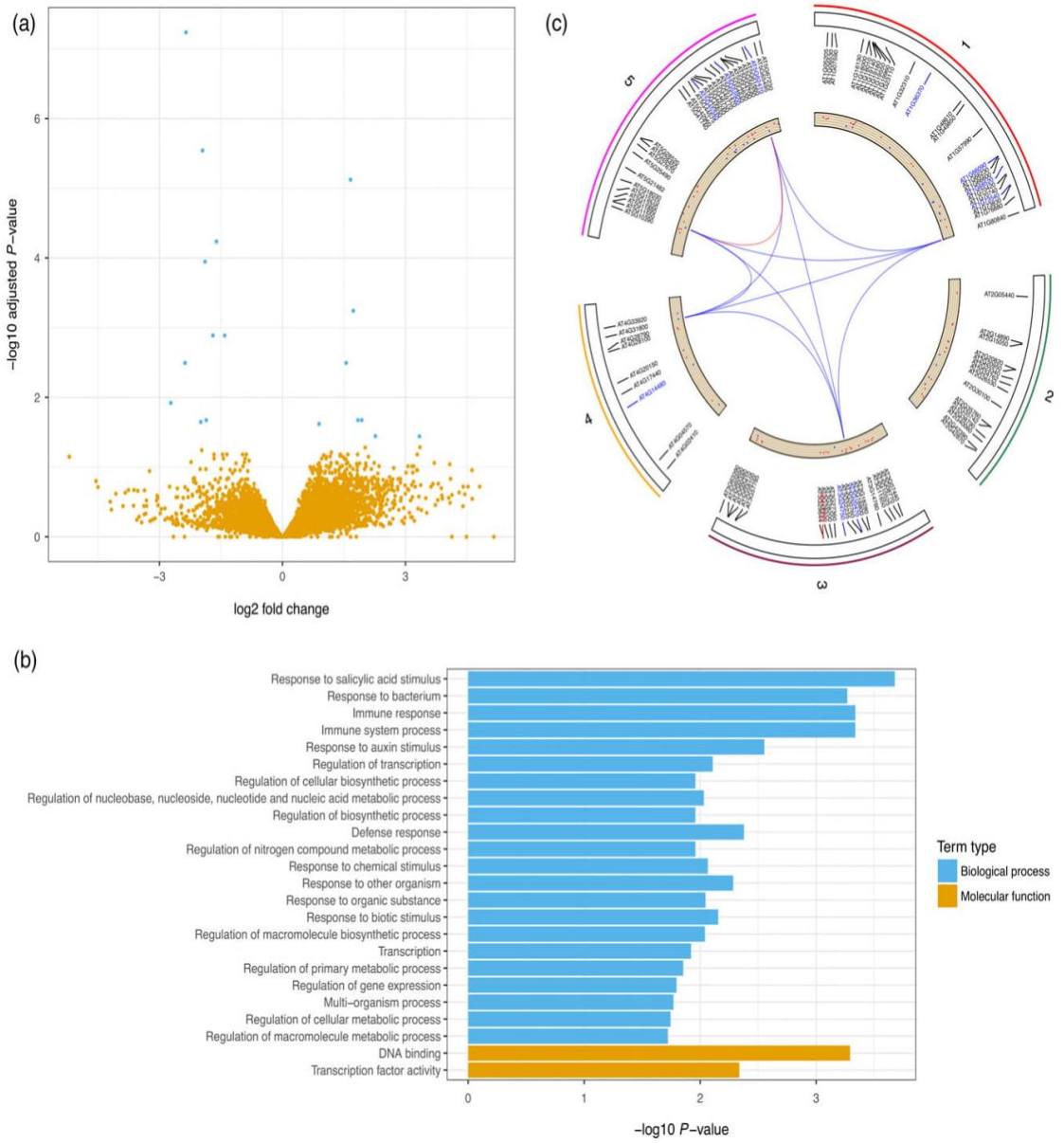


Figure 7. Gene differential expression analysis of *TE-linc11195* mutant using RNA-seq.
(a) Volcano plot showing log2 fold changes versus statistical significances of genes. Blue dots represent statistically significant differentially expressed genes (Benjamini–Hochberg method adjusted *P*-value < 0.05).
(b) GO enrichment analysis of 100 most significantly differentially expressed genes.
(c) Genomic distribution of 100 most significantly differentially expressed genes. Gene labels with blue colour are top 10 most significantly expressed genes. Scatter plot inside inner track represents log2-fold changes of genes, therefore, red and blue dots represent up- and down- regulated genes respectively. Links inside circle plot represent five genes associated with most significant over-represented GO term 'response to salicylic acid stimulus', blue and red lines represent between- and in- chromosome connections respectively. [Colour figure can be viewed at wileyonlinelibrary.com].

observed in plants defected in TE-linc11195 at stages of seed germination and post-germination seedling development (Figure S11b, c). These results further indicate that TE-lincRNAs are involved in plants' responses to abiotic stress. Because it has been suggested that TEs provide unique sequence elements to conserved lncRNAs (Hezroni *et al.*, 2015), the contribution of the LTR to TE-linc11195 was also checked under ABA treatment. Interestingly, we found that this LTR could strengthen the extent of ABA response for TE-linc11195 (Figure S12), indicating that TEs play a biological role in the evolution of lincRNAs.

Changes in chromatin state caused the generation of unique lincRNAs in *ddm1* Col (Table 1 and Figure 8). These unique lincRNAs may also play a role in responses to stress, which may contribute to the biotic stress resistance found in *ddm1* Col (Dowen *et al.*, 2012). Our observation is also different from the previous suggestion that TE insertions give rise to functional lncRNAs (Ponting

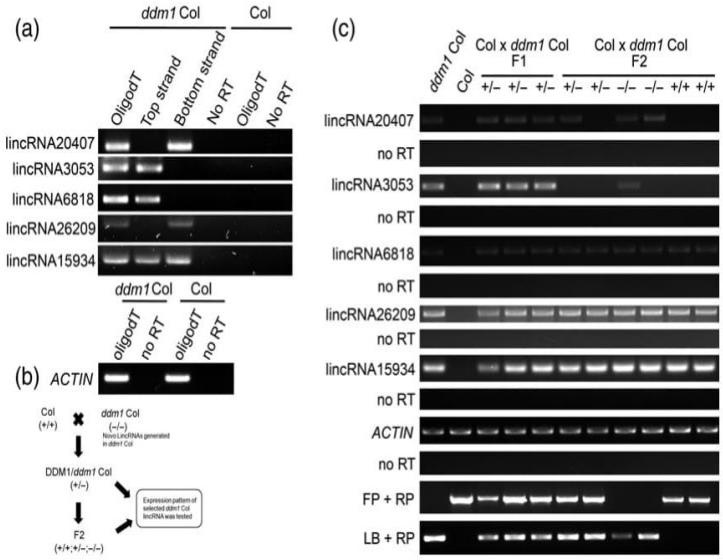


Figure 8. Characterization of unique lincRNAs generated by loss of DDM1.
(a) Strand-specific RT-PCR analysis was performed on selected lincRNAs only present in the *ddm1* mutant, three TE-lincRNAs: lincRNA20407, lincRNA3053 and lincRNA6818; two non-TE-lincRNAs: lincRNA26209 and lincRNA159.
(b, c) Expression pattern of *ddm1* dependent lincRNAs in subsequent generations. The – or + symbol indicates the presence or absence of the mutant or wild-type DDM1 allele, respectively. Actin was used as a positive control. FP, RP and LB are primers used to genotype the plants. Primers LB and RP indicate the presence of the *ddm1* T-DNA and primers FP and RP indicate the presence of wild-type allele.

et al., 2009), indicating that both TE-lincRNA and non-TE-lincRNA can simply arise by alteration of chromatin state. This finding provides an attractive hypothesis that chromatin altered by environmental factors can produce unique lincRNAs which may be functional when responding to the environment and can be inherited. Our hypothesis is also consistent with a previous suggestion that lncRNAs have a distinct advantage over proteins as gene regulators because they can be functional immediately upon transcription without needing to be translated into protein outside the nucleus (Johnson and Guigo, 2014). In the light of the many possible regulatory roles of lincRNAs, the environmentally triggered appearance of lincRNAs may diversify biological regulation of the organism and drive an increased rate of evolution. Our observation that TE-lincRNA11195 was transcribed in the genus Arabidopsis but not *Capsella* (Figure S13) might help explain lineage-specific changes in gene networks. As transposable elements are often clade specific, clade specific TE-lincRNAs would be expected to frequently arise. This idea could be tested by RNA-seq analysis to identify lineage-specific TE-lincRNAs from a number lineages combined with CRISPR/Cas genome editing to remove specific lineages of TE-lincRNAs.

CONCLUSION

We have identified 47, 611 and 398 TE-linRNAs in 2-week-old seedlings of *Arabidopsis thaliana*, rice and maize respectively. Different TE families have differing extents of contribution to lincRNAs. More importantly, we found that many TE-lincRNAs are potentially stress-responsive and may contribute to stress response. This was validated by the perturbation of one TE-lincRNA, lincRNA11195, which was found to be involved in the ABA response. Furthermore, unique TE-lincRNAs and non-TE-lincRNAs could be detected in mutants whose nuclear chromatin state had changed, and these unique lincRNAs were inherited. This research has evaluated the contribution of TEs to lincRNAs and demonstrated the important role played by TE-lincRNAs in response to stress.

EXPERIMENTAL PROCEDURES

RNA-seq library preparation and sequencing

Total RNAs were obtained from 2-week-old seedlings of Arabidopsis, rice and maize. The preparation of strand-specific RNA-seq libraries and deep sequencing were performed in the Shanghai Center for Plant Stress Biology (Shanghai, China). These libraries were constructed through applying TruSeq Stranded mRNA (Illumina, San Diego, CA, USA) in accordance with the manufacturer's instruction. The quality of RNA-seq libraries were assessed by using a Fragment Analyzer (Advanced Analytical, IA, USA), and the resulting libraries were sequenced on an Illumina HiSeq 2500 instrument producing pair-end reads of 100 or 125 nucleotides. For *ddm1* Col, RNA was extracted from 2-week-old seedlings, and shipped to Beijing Genomics Institue (Shenzhen, China) for sequencing.

TE-lincRNA identification pipeline

Adaptors and low quality sequences were filtered with trim-galore (v0.3.3, –stringency 6). Then clean reads were aligned to reference genomes (TAIR10 for Arabidopsis, TIGR release 7 for

rice and AGPv2 for maize) using Tophat2 with following parameters: -N 5 –read-edit-dist 5 (v2.0.14) (Kim *et al.*, 2013). Mapped reads from three biological replicates for Arabidopsis and rice were merged and then assembled with Cufflinks respectively (v2.2.1) (Trapnell *et al.*, 2010). For maize, mapped reads were assembled with Cufflinks firstly and then merged with Cuffmerge, due to the large number of mapped reads (Trapnell *et al.*, 2010). Annotated protein-coding genes or transcripts with protein encoding potential were filtered with following three steps: (i) remove short transcripts (shorter than 200 bp), intronic transcripts and transcripts overlapping with protein-coding genes (at least 1 bp overlapping); (ii) BLASTX against SWISSPROT protein sequence database (Camacho *et al.*, 2009); and (iii) remove transcripts with ORFs longer than 100 aa inside or 50 aa at end(s). The remaining transcripts were categorized as lincRNAs. Finally, genomic coordinates of lincRNAs were further checked with respect to TEs in Arabidopsis, rice and maize respectively. LincRNAs overlapping with but not fully inside TE (s) were characterised as TE-lincRNAs.

Sequence conservation analysis

Whole-genome level pairwise alignments of Arabidopsis with 23 other plants and rice with 27 other plants were downloaded from Ensemble Plants (Kersey *et al.*, 2012). Multiple alignments were obtained by merging pairwise alignments with multiz (Blanchette *et al.*, 2004). Phylogenetic models were estimated by applying phyloFit on four-fold degenerate (4d) sites according to the manual (Hubisz *et al.*, 2011). Based on the multiple alignments and estimated phylogenetic models, conservation scores for different genomic features, including protein-coding genes, TEs, TE-lincRNAs, non-TE-lincRNAs and intergenic intervals (the intergenic intervals were defined as the genomic intervals after removing all protein-coding genes and lincRNAs), were calculated by using phyloP with following parameters: –features –method SCORE –mode CONACC (Hubisz *et al.*, 2011).

RNA motif detection

Rfam 12.0 is a collection of noncoding RNA families by multiple sequence alignments, consensus secondary structures and covariance models (CMs) (Nawrocki *et al.*, 2015). The program 'cmscan' from the infernal package was used to search the lincRNA sequence against CM-format motifs in Rfam 12.0 with following parameter: –E $1e^{-1}$ (Nawrocki and Eddy, 2013). If multiple RNA motifs were identified from overlapped regions the one with the smallest E-value was selected.

Expression correlation analysis and co-expression network reconstruction

Variance-stabilizing transformation of raw counts for lincRNAs and protein-coding genes across multiple samples from public RNA-seq datasets (SRA00903 and GSE49325) were used to calculate pairwise correlation between transcripts. Pearson's correlation was calculated between lincRNA and the 10 closest protein-coding genes. WGCNA was used to reconstruct Arabidopsis lincRNA and reference gene co-expression networks (Langfelder and Horvath, 2008).

Statistical analysis and data visualization

Statistical analysis and data visualization of characterises of TE-linRNAs and non-TE-lincRNAs were performed with R and R packages (Lawrence *et al.*, 2009; R Development Core Team, 2010, Yin *et al.*, 2012).

Plant materials, stress treatment and PCR assay

Seeds of *C. rubella* and *A. thaliana* T-DNA insertion mutants including *11195-1* (CS843057), *11195-2* (CS834193) and *ddm1-10* (SALK_093009) were obtained from Arabidopsis Biological Resource Center (ABRC). ABA insensitive mutant used in this study is *pyr1/pyl1/pyl4* (Park *et al.*, 2009). For generating transgenic lincRNA11195 plants with or without the LTR, DNA fragments containing 1.5 kb upstream of lincRNA11195 and the full-length or lacking LTR region lincRNA sequence plus a 200-bp downstream sequence with attB sites were amplified from Col-0 genomic DNA, and were then cloned into Gateway vector pDONR207 (Invitrogen). Each insert was subsequently introduced into the Gateway pGWB1 vector by LR reaction (Invitrogen). All plasmids were transformed into *Agrobacterium tumefaciens* strain GV3101, and then transformed into *A. thaliana* plants of the mutant backgrounds via the floral dip method. Stress treatment was carried out as described previously (Zeller *et al.*, 2009). Preparation of cDNA and real-time quantitative PCR were performed according to the previous description (Wang *et al.*, 2014). RT-PCR and strand-specific RT-PCR were carried out as described previously (Wierzbicki *et al.*, 2008). All experiments were carried out with at least three biological replicates. Details of the primers used in this study are listed in Table S8.

Gene differential expression analysis of *TE-lincRNA11195* mutant RNA-seq

Fourteen-day-old wild-type and *11195-2* seedlings were grown on half-strength MS medium then treated with either 0 or 100 μM ABA for 12 h, RNA extracted and then Illumina sequencing performed. Adaptor and low quality sequences were trimmed with trim_galore the same as above. Clean reads were aligned to reference genome using STAR_2.5.2a with following parameters: –outFilterMismatchNmax 10 –outFilterMismatchNoverLmax 0.05 –seedSearchStartLmax 30. Gene differential expression analysis was performed using edgeR with GLM method considering two factors: lincRNA11195 mutant and ABA treatment (Robinson *et al.*, 2010).

Sequence pairwise alignment and secondary structure prediction of TE-lincRNA11195 in *A. thaliana* and *A. lyrata*

Homolog of TE-lincRNA11195 in *A. lyrata* was determined using its sequence of *A. thaliana* blastn against *A. lyrata* genomic sequences (https://github.com/PacificBiosciences/DevNet/wiki/Arabidopsis-lyrata) and extended to the equivalent length of TE-linc11195 in *A. thaliana*. Sequence pairwise alignment of TE-lincRNA11195 between *A. thaliana* and *A. lyrata* was performed using ClustalX2 (Larkin *et al.*, 2007). The secondary structures of TE-lincRNA11195 in two species were predicted using RNAfold with the default setting (Gruber *et al.*, 2008).

Availability of data and materials

The data sets supporting the results of this article are available in NCBI's GEO database repository, and are accessible through GEO accession number GSE76798.

AUTHORS' CONTRIBUTIONS

Experiments were designed by DW, ZQ, QZ, DA, IS and JK. DW, ZQ, LY, QZ, ZL, TD, and ZW conducted experiments

and all authors analysed the data. DW, ZQ, IS wrote the manuscript and all authors edited the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

This research was funded by the Chinese Academy of Sciences, and National Science Foundation of China (31401077) awarded to DW, by the Australian Research Council (ARC) through a Future Fellowship (FT130100525) awarded to IS and a MOET-VIED PhD scholarship awarded to TD. The authors declare no conflicts of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Nuclear distribution of TE-lincRNAs and non-TE-lincRNAs in Arabidopsis (a), rice (b), and maize (c).

Figure S2. Length distribution of TE-lincRNAs and non-TE-lincRNAs identified from Arabidopsis (a), rice (b) and maize (c).

Figure S3. Exon numbers of TE-lincRNAs and non-TE-lincRNAs identified from Arabidopsis (a), rice (b) and maize (c).

Figure S4. Numbers of RNA motifs detected in TE-lincRNAs and non-TE-lincRNAs from Arabidopsis (a), rice (b) and maize (c).

Figure S5. Distribution of the distance of lincRNAs to their corresponding nearby genes in Arabidopsis (a), rice (b) and maize (c).

Figure S6. Correlation of expression between lincRNAs and 10 closest genes and the example of lincRNA (a) and protein-coding gene co-expression network showing stress response (b, c).

Figure S7. Contribution of TEs to lincRNAs in Arabidopsis, rice and maize.

Figure S8. Comparison of lincRNAs and protein-coding genes contributed by TEs in Arabidopsis, rice and maize.

Figure S9. Mutated TE-lincRNA11195 alters sensitivity to ABA in seed germination and post-germination development in Arabidopsis.

Figure S10. Expression of Arabidopsis TE-lincRNA11195 in WT and an ABA-insensitive mutant *pyr1/pyl1/pyl4*.

Figure S11. Mutated TE-lincRNA11195 alters sensitivity to salt in seed germination and post-germination seedling development in Arabidopsis.

Figure S12. TE strengthens the ABA-responsive transcription to TE-lincRNA11195 in Arabidopsis.

Figure S13. TE-lincRNA11195 is detected only in Arabidopsis genus.

Figure S14. Sequence pairwise alignment and secondary structure prediction of TE-lincRNA11195 in *A. thaliana* and *A. lyrata*.

Table S1. General information about RNA sequence libraries used in this study.

Table S2. Genomic coordinates of TE-lincRNAs in three species.

Table S3. Summary of Rfam RNA motifs detected in TE-lincRNAs and non-TE-lincRNAs from three species.

Table S4. Summary of co-expression network analysis in Arabidopsis.

Table S5. Statistics of different TE families contributing to lincRNAs in three species.

Table S6. Summary of differential gene expression between wild-type and mutant *TE-linc11195-2*.

Table S7. Non-Mendelian inheritance of *ddm1* induced lincRNAs.

Table S8. Primers used in this study.

REFERENCES

Alonso, J.M., Stepanova, A.N., Leisse, T.J. *et al.* (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, **301**, 653–657.

Ariel, F., Jegu, T., Latrasse, D., Romero-Barrios, N., Christ, A., Benhamed, M. and Crespi, M. (2014) Noncoding transcription by alternative RNA polymerases dynamically regulates an auxin-driven chromatin loop. *Mol. Cell*, **55**, 383–396.

Ariel, F., Romero-Barrios, N., Jegu, T., Benhamed, M. and Crespi, M. (2015) Battles and hijacks: noncoding transcription in plants. *Trends Plant Sci.* **20**, 362–371.

Bennetzen, J.L. and Wang, H. (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* **65**, 505–530.

Blanchette, M., Kent, W.J., Riemer, C. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715.

Bologna, N.G. and Voinnet, O. (2014) The diversity, biogenesis, and activities of endogenous silencing small RNAs in Arabidopsis. *Annu. Rev. Plant Biol.* **65**, 473–503.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Cech, T.R. and Steitz, J.A. (2014) The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*, **157**, 77–94.

Derrien, T., Johnson, R., Bussotti, G. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789.

Di, C., Yuan, J., Wu, Y. *et al.* (2014) Characterization of stress-responsive lncRNAs in Arabidopsis thaliana by integrating expression, epigenetic and structural features. *Plant J.* **80**, 848–861.

Dowen, R.H., Pelizzola, M., Schmitz, R.J., Lister, R., Dowen, J.M., Nery, J.R., Dixon, J.E. and Ecker, J.R. (2012) Widespread dynamic DNA methylation in response to biotic stress. *Proc. Natl Acad. Sci. USA*, **109**, E2183–E2191.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.K. and Mockler, T.C. (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res.* **20**, 45–58.

Fort, A., Hashimoto, K., Yamada, D. *et al.* (2014) Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.* **46**, 558–566.

Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I., Puga, M.I., Rubio-Somoza, I., Leyva, A., Weigel, D., Garcia, J.A. and Paz-Ares, J. (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat. Genet.* **39**, 1033–1037.

Goff, L.A., Groff, A.F., Sauvageau, M. *et al.* (2015) Spatiotemporal expression and transcriptional perturbations by long noncoding RNAs in the mouse brain. *Proc. Natl Acad. Sci. USA*, **112**, 6855–6862.

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neubock, R. and Hofacker, I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Res.* **36**, W70–W74.

Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P. and Ulitsky, I. (2015) Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* **11**, 1110–1122.

Hubisz, M.J., Pollard, K.S. and Siepel, A. (2011) PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51.

Ito, H., Gaubert, H., Bucher, E., Mirouze, M., Vaillant, I. and Paszkowski, J. (2011) An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature*, **472**, 115–119.

Johnson, R. and Guigo, R. (2014) The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA*, **20**, 959–976.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M. and Feschotte, C. (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9**, e1003470.

Kelley, D. and Rinn, J. (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* **13**, R107.

Kersey, P.J., Staines, D.M., Lawson, D. *et al.* (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.* **40**, D91–D97.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.

Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Larkin, M.A., Blackshields, G., Brown, N.P. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Lawrence, M., Gentleman, R. and Carey, V. (2009) rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, **25**, 1841–1842.

Li, L., Eichten, S.R., Shimizu, R. *et al.* (2014) Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biol.* **15**, R40.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C. and Chua, N.H. (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell*, **24**, 4333–4345.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., Okumoto, Y., Tanisaka, T. and Wessler, S.R. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*, **461**, 1130–1134.

Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Nawrocki, E.P., Burge, S.W., Bateman, A. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–D137.

Park, S.Y., Fung, P., Nishimura, N. *et al.* (2009) Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. *Science*, **324**, 1068–1071.

Plath, K., Fang, J., Mlynarczyk-Evans, S.K., Cao, R., Worringer, K.A., Wang, H., de la Cruz, C.C., Otte, A.P., Panning, B. and Zhang, Y. (2003) Role of histone H3 lysine 27 methylation in X inactivation. *Science*, **300**, 131–135.

Ponting, C.P., Oliver, P.L. and Reik, W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.

R Development Core Team. (2010) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Talbert, P.B. and Henikoff, S. (2014) Environmental responses mediated by histone variants. *Trends Cell Biol.* **24**, 642–650.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515.

Ulitsky, I. and Bartel, D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell*, **154**, 26–46.

Wang, D. and Timmis, J.N. (2013) Cytoplasmic organelle DNA preferentially inserts into open chromatin. *Genome Biol. Evol.* **5**, 1060–1064.

Wang, D., Qu, Z., Adelson, D.L., Zhu, J.K. and Timmis, J.N. (2014) Transcription of nuclear organellar DNA in a model plant system. *Genome Biol. Evol.* **6**, 1327–1334.

Wang, M., Yuan, D., Tu, L., Gao, W., He, Y., Hu, H., Wang, P., Liu, N., Lindsey, K. and Zhang, X. (2015) Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.). *New Phytol.* **207**, 1181–1197.

Wang, X., Ai, G., Zhang, C., Cui, L., Wang, J., Li, H., Zhang, J. and Ye, Z. (2016) Expression and diversification analysis reveals transposable elements play important roles in the origin of Lycopersicon-specific lncRNAs in tomato. *New Phytol.* **209**, 1442–1455.

Wicker, T., Sabot, F., Hua-Van, A. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982.

Wierzbicki, A.T., Haag, J.R. and Pikaard, C.S. (2008) Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell*, **135**, 635–648.

Yin, T., Cook, D. and Lawrence, M. (2012) ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.* **13**, R77.

Zeller, G., Henz, S.R., Widmer, C.K., Sachsenberg, T., Ratsch, G., Weigel, D. and Laubinger, S. (2009) Stress-induced changes in the Arabidopsis thaliana transcriptome analyzed using whole-genome tiling arrays. *Plant J.* **58**, 1068–1082.

Zhang, W., Zhang, T., Wu, Y. and Jiang, J. (2012) Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *Plant Cell*, **24**, 2719–2731.

Zhang, Y.C., Liao, J.Y., Li, Z.Y., Yu, Y., Zhang, J.P., Li, Q.F., Qu, L.H., Shu, W.S. and Chen, Y.Q. (2014) Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. *Genome Biol.* **15**, 512.

Zhu, Q.H., Stephen, S., Taylor, J., Helliwell, C.A. and Wang, M.B. (2014) Long noncoding RNAs responsive to Fusarium oxysporum infection in Arabidopsis thaliana. *New Phytol.* **201**, 574–584.

Chapter 5: Identification of PRC2-associated Long Noncoding RNA in *Arabidopsis thaliana*

Trung Do[1], Zhipeng Qu[1], Jun Li[1], Rakesh David[1], Dave Adelson[1], Chris Helliwell[2] and Iain Searle[1,*]

[1]Department of Molecular and Biomedical Sciences, School of Biological Sciences, The University of Adelaide, Adelaide, South Australia 5005, Australia.

[2]Commonwealth Scientific Industrial Research Organization (CSIRO), CSIRO Agriculture and Food, Canberra, Australia.

[*] Corresponding author: iain.searle@adelaide.edu.au

In preparation for BMC Plant Biology

Statement of Authorship

Title of paper	Identification of PRC2-associated long non-coding RNA in *Arabidopsis thaliana*
Publication Status	In preparation
Publication details	To be submitted to BMC Plant Biology.

Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. The candidate's stated contribution to the publication is accurate (as detailed below)
ii. Permission is granted for the candidate in include the publication in the thesis and
iii. The sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Candidate	Trung Q. Do		
Contribution to paper	Co-Designed experiments. Wrote, edited the manuscript and composed figures.		
Overall percentage (%)	70%		
Signature		Date	29/05/18

Name of Co-Author	Zhipeng Qu		
Contribution to paper	Bioinformatic analysis of RNA-seq data. Wrote the Materials and Methods section on bioinformatics analysis. Overall contribution 10%.		
Signature		Date	30/05/18

Name of Co-Author	Jun Li		
Contribution to paper	Bioinformatic analysis of RNA-seq data. Overall contribution 5%.		
Signature		Date	30/5/2018

Name of Co-Author	Rakesh David		
Contribution to paper	Contributed to method development. Overall contribution 2%.		
Signature		Date	1/6/18

Name of Co-Author	David L. Adelson		
Contribution to paper	Supervised development of bioinformatics analysis. Overall contribution 2%.		
Signature		Date	1/6/18

Name of Co-Author	Chris Helliwell		
Contribution to paper	Co-designed experiments and interpreted results. Overall contribution 2%.		
Signature		Date	4/6/18

Name of Co-Author	Iain R. Searle		
Contribution to paper	Co-designed experiments. Supervised development of work and edited the manuscript. Overall contribution 9%.		
Signature		Date	29/5/18

Abstract

Background

Polycomb group (PcG) complexes form evolutionarily conserved multi-protein complexes that play critical roles in the control of developmental processes in plants and other eukaryotes. In *Arabidopsis thaliana*, the PcG repressive complex 2 (PRC2) proteins are grouped into three distinct complexes, EMF2–PRC2 (EMF2; EMBRYONIC FLOWER 2), VRN2–PRC2 (VRN2; VERNALIZATION 2) and FIS2–PRC2 (FIS2; FERTILIZATION-INDEPENDENT SEED 2). FIS2–PRC2 is restricted to the female gametophyte and seed tissues and is essential for normal seed development.

Results

We immunoprecipitated FIS2 from endosperm tissue of developing *Arabidopsis* seeds and sequenced the associated RNAs by using Illumina sequencing. We identified 16,637 associated long noncoding RNAs (lncRNAs). The identified lncRNAs showed shorter transcripts, lower expression levels and more specific expression than did protein-coding genes. Additionally, with the aim of identifying potential regulatory target genes of PRC2-associated lncRNAs, the expression correlation between PRC2-associated lncRNAs and the upregulated protein-coding genes from the *fis2* mutant transcriptome was assessed. We identified both positive and negative correlations. Importantly, G-tract motifs were significantly enriched among PRC2-binding transcripts.

Conclusion

Thousands of lncRNAs are bound to the FIS2–PRC2 complex through the G-tract motif to regulate gene expression in *A. thaliana*. In future, these PRC2-associated lncRNAs would be beneficial for understanding the variation in gene regulation by FIS2–PRC2 complex in plants.

Keywords: *Arabidopsis thaliana*, FIS2, G-tract motifs, H3K27me3, long noncoding RNA, PRC2

Introduction

Transcriptome analysis of the mammalian genome has shown that although the proportion of protein-coding genes is only 1–2%, 70–90% of genes are transcribed from various regions such as intergenic or intronic regions [1–3]. This means that most of those transcripts (from 100 nt to >10 kb) are noncoding and their functions are unknown [2]. Long noncoding RNAs (lncRNAs) have a potential role in many regulatory processes in eukaryotes. Most lncRNAs participate in gene regulation by modulating transcriptional activity through the interaction with regulatory protein complexes such as the polycomb group repressive complex 2 (PRC2) complex [4–5]. These interactions regulate epigenetic changes at the target site [6–7]. However, the molecular action of lncRNAs in this context is not well understood.

In mammals, the histone methyltransferase PRC2 is a multiple complex. It is made up of multiple protein (e.g. EZH2, SUZ12, EED, RBBP4 and JARID2) and is required for various epigenetic silencing processes during embryonic development and cancer cell growth [8]. In *Drosophila,* with the help of several co-factors, PRC2 is recruited to chromatin through binding to a polycomb response element (PRE) [9–10]. However, the exact mechanism for PRC2 recruitment is not clear because no PRE-like elements have been reported in mammals [10] and the process may depend on assembling factors such as DNA elements, bridging proteins and lncRNAs [9–10]. Studies have demonstrated roles of lncRNAs in recruiting PRC2 by binding to this complex and guiding them to the target sites [11–15]. For example, in humans, the antisense lncRNA *HOTAIR*, transcribed from the HOXC locus, associates physically with the PRC2 complex, modulating PRC2 activity to deposit trimethylated lysine 27 on histone H3 (H3K27me3) marks at the HOXD locus [12–13]. In animals, the RNA–PRC2 interaction has been studied *in vitro* and *in vivo* [7, 9, 16]. From electrophoretic mobility shift assays and RNA pull-down experiments, the authors have shown that PRC2 proteins bind to RepA RNA more specifically than they do to non-relevant RNA transcripts [7, 9] and *cis*-acting RNAs block the histone methyltransferase activity of PRC2 until the RNA–PRC2 complex combines with JARID2 [17]. Although the molecular nature of the interaction between lncRNAs

and PRC2 is yet to be determined, the interaction between lncRNAs and chromatin-modifying complexes appears to represent a general mechanism for epigenetic repression in animals.

In plants, there are few reports of the functions of lncRNAs. The first intergenic lncRNAs to be induced by phosphate starvation was discovered in *Medicago truncatula* (*Mt4*), *A. thaliana* (*IPS1*, *INDUCED BY PHOSPHATE STARVATION1* and *At4*), tomato (*Lycopersicon esculentum* L.; *TPSI1*, *TOMATO PHOSPHATE STARVATION-INDUCED GENE 1*) and rice (*Oryza sativa*; *OsPI1*, *ORYZA SATIVA PHOSPHATE-LIMITATION INDUCIBLE GENE 1*) [18–21]. Another intergenic lncRNA had been reported to function during pollen development in rice under long-day conditions [22]. In *Arabidopsis*, two lncRNAs (*COLD-INDUCED LONG ANTISENSE INTRAGENIC RNA* [*COOLAIR*] and *COLD-ASSISTED INTRONIC NONCODING RNA* [*COLDAIR],* have been shown to interact with CURLY LEAF (CLF) of PRC2 during vernalisation to control *FLOWERING LOCUS C* (*FLC*) by promoting methylation [5–6, 23]. Recently, a number of lncRNAs have been shown to be differentially expressed in response to stress stimuli in *Arabidopsis* [24-25] and rice [26]. All of these reports provide evidence for the prominent role of lncRNAs in the regulation of plant growth, development and stress responses.

Molecular studies have shown that *COOLAIR* and *COLDAIR* lncRNAs in plants play similar roles to those of *HOTAIR* and *Xist* noncoding RNA in animals in acting to recruit PRC2 complex to target chromatin [27]. These data together suggest that lncRNA-mediated epigenetic gene silencing by PRC2 complex may be an evolutionarily conserved mechanism in plants and animals. This interaction plays an important role in plant development; thus, understanding its molecular mechanisms will enhance our efforts in plant breeding and regulation of plant development.

Of significant interest is the *A. thaliana* silique, which is developed from the ovule of the flower. Towards the identification of molecular mechanisms of endosperm development, we generated comprehensive RNA-seq datasets from 1DAP (1 day after pollination) siliques of HA-tagged-FIS2 transgenic lines to profile genome-wide expression of PRC2-associated lncRNAs. In the current work, we examine

lncRNAs from transcriptomes of 1DAP siliques of HA-tagged-FIS2 transgenic lines with the aim of finding PRC2-associated lncRNAs that function in endosperm development. In total, we identified 16,637 lncRNAs from the PRC2-associated lncRNA transcriptome datasets. The transcriptome analysis also showed that these lncRNAs have gene structure and transcription regulation that is similar to that of protein-coding genes. However, they also have some distinct features, such as (1) a large number of lncRNAs are from a single exon; (2) they are expressed at a low level (reads per kilobase million (RPKM) ~1); and (3) they are small in length (200–500 bp). With the aim of identifying potential regulatory target genes of PRC2-associated lncRNAs, the expression correlation between PRC2-associated lncRNAs and the upregulated protein-coding genes from the *fis2* mutant transcriptome was assessed. We identified both positive and negative correlations. Our analysis supports observations of the PRC2-associated lncRNAs landscape in seed development and provides a foundation for future research into the function of PRC2-associated lncRNAs in *A. thaliana.*

Materials and methods

Plant materials

A. thaliana (Columbia-0 accession) wild type and transgenic lines were grown in Phoenix Biosystems growth under metal halide lights as described previously [28]. For plate experiments, seeds were surface sterilized for 12 hours using chlorine gas, plated on ½ MS medium supplemented with 1% sucrose and sealed as described previously [29]. All plants were grown under long-day photoperiod conditions of 16-h light and 8-h darkness at 21^{o}C.

Homozygous FIS2 promoter::FIS2:HA (pFIS2::FIS2:HA) epitope-tagged transgenic lines were constructed by Chris Helliwell (The Australian National University) and transformed into Columbia wild type. Transgenic plants were selected on ½ MS media supplemented with 15 μg ml^{-1} Hygromycin B. FIS2 transcript abundance was assessed in at least five independent T_1 plants using quantitative reverse transcription PCR (RT-qPCR) and two lines with FIS2 mRNA abundance similar to wild type transcript levels were carried through to homozygous T_3 generation for molecular analysis. Siliques from pFIS2::FIS2:HA

or wild type were hand pollinated and harvested 1 day later for RNA immunoprecipitation (RIP) experiments.

RIP and RIP-seq

Siliques from two biological replicates of either wild type or pFIS2::FIS2:HA 1DAP siliques were snap frozen in liquid nitrogen and used immediately or stored at –80^{o}C. One g of siliques was ground to a fine powder using a mortar and pestle and RIP performed following the protocol described by Köster and Staiger [30] with several modifications. Immunoprecipitation (IP) was performed with an anti-HA antibody (1:10,000, abcam, product code ab9110) with or without formaldehyde cross-linking of the ground tissue.

To construct RIP-seq libraries, whole cell lysates were prepared from formaldehyde-fixed siliques, treated with 400 U ml^{-1} DNAse I (New England Biolabs) and 20 U ml^{-1} RNaseOUT™ (ThermoFisher Scientific), and incubated with anti-HA antibodies beads (ThermoFisher Scientific) for 2 h. Total RNA was extracted using TRIzol (ThermoFisher Scientific). Total RNA-seq libraries were then constructed using a NEBNext Ultra Directional RNA library Prep Kit (New England Biolabs) and sequenced using an Illumina HiSeq™ 2500 in paired-end (PE 100) mode.

Transcriptome detection by RNA-seq

Adaptor and low-quality sequences of raw reads were removed using trim_galore with the following parameters: stringency 6, with trimmed reads then aligned against the *A. thaliana* TAIR10 genome assembly using *TopHat2* with parameters -N 5 -- read-edit-dist 5. Aligned reads from all samples were merged with SAMtools and assembled using *Cufflinks* with the following parameters: --library-type fr-firststrand -u. Assembled transcripts were filtered through our lncRNA identification pipeline as described previously [25]. Transcripts shorter than 200 nt were removed and genomic coordinates of long transcripts were checked against reference genes of TAIR10 and classified as either gene transcripts, intergenic transcripts, intronic transcripts or antisense transcripts. The latter three classes of transcript were selected to filter unannotated protein-coding potential transcripts by following two steps: 1) a sequence similarity search

against the Swiss-Prot protein database; and 2) prediction of open reading frame(s) (ORFs).

Calculation of lncRNA conservation

To calculate the conservation of the RIP *A. thaliana* lncRNAs, datasets for lncRNAs from other Brassicaceae plants including *Brassica rapa*, *B. napus* and *B. oleracae* were collected from CANTATAdb [31]; their genomes were downloaded and then aligned with the sequences of FIS2-associated lncRNAs using the BLASTN 2.6.0+ software on the NCBI website (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

RT-qPCR

Validation RIPs were performed as described [26]. RIP was followed by quantitative, strand-specific RT-PCR. First-strand cDNA synthesis was carried out using approximately 300 ng RNA and SuperScript™ III (ThermoFisher Scientific). All primers used in this study are listed in Table S1 (see Appendices). RT-qPCR was performed in quadruplicate using the SYBR Green Mastermix (Roche Applied Science) on a Roche Light Cycler 480 (Roche Applied Science) according to the manufacturer's instructions. Sample cycle threshold (Ct) values were determined and standardized relative to the input, and the $2^{-\Delta CT}$ method was used to calculate the relative changes in gene expression based on the RT-qPCR data.

Identification of G-quadruplex-forming sequences (GQSes)

Whole sequences of 16,637 PRC2-associated lncRNAs identified from *A. thaliana* early development siliques were used. These sequences were scanned using the Quadparser tool [32] for GxNy1GxNy2GxNy3Gx, where x = G2 or G3; y = 1/1–2/1–4 for G2 and 1–3/1–7 for G3. The different categories were defined as follows: loop 1–3, $(G3N\{1–3\})_3G3$ with N = [ATCG], loop 1–7, $(G3N\{1–7\})_3G3$ and loop 1, $(G2N\{1\})_3G2$, loop 1–2 $(G2N\{1–2\})_3G2$ and loop 1–4, $(G2N\{1–4\})_3G2$.

Results

Identification of FIS2–PRC2-associated lncRNAs from *Arabidopsis* endosperm

In both the plant and animal kingdoms, the evolutionarily conserved polycomb-mediated gene repression and maintenance is important for cell identity and developmental processes [27, 33]. In animals, the PRC2 complex has been shown to interact with a large number of RNA transcripts [33] but little is known about the bound RNA transcripts and their functional role in plants. To address this important knowledge gap, we developed an IP protocol for the FIS2–PRC2 complex from developing *A. thaliana* endosperm tissue and sequenced the FIS2-bound RNAs.

FIS2 expression is restricted to the female gametophyte and developing endosperm tissue in *Arabidopsis* [34]; therefore IP of FIS2 from whole siliques will lead to isolation of FIS2 from endosperm tissue. To identify FIS2–PRC2-associated RNAs, we produced a single-insert, epitope-tagged pFIS2::FIS2:HA transgenic line and developed a stringent IP protocol using an anti-HA antibody. Briefly, we developed an IP protocol such that after IP and stringent washing from wild type siliques, no RNA was detected using a Bioanalyzer RNA pico chip (data not shown). Given that very small amounts (less than 50 pg) of RNA may have been present, we attempted to construct an Illumina library; after quality control using a Bioanalyzer DNA chip we detected no inserts in the library—only an adapter–adapter band (data not shown). In contrast, our positive control inserts were successfully cloned (data not shown). Therefore, we concluded that by using our stringent IP protocol, only FIS2 should be immunoprecipitated from pFIS2::FIS2:HA tissue and the FIS2-associated RNAs sequenced.

We harvested biological replicates of silique tissue from transgenic plants, immunoprecipitated FIS2, purified the associated RNAs, constructed libraries and Illumina sequenced them. The bioinformatic pipeline to analyze the sequencing data and identify lncRNAs is outlined in Figure 1A. Briefly, sequence reads were aligned using *TopHat2* to preserve junction reads (Table 1), then *Cufflinks* was used to assemble uniquely mapped reads into known and novel transcripts; these transcripts were combined by *Cuffmerge* and then compared with the reference annotation by using *Cuffcompare*. Based on the availability of

very large deep sequencing datasets, the merged transcripts were compared with known protein-coding genes and lncRNAs in public databases to obtain a minimum number of novel, false positive lncRNA transcripts (Figure 1B). We identified 55,627 reliably expressed transcripts and among those, 55,317 were longer than 200 nt. Of these 55,317 transcripts, 16,637 long transcripts were identified as lncRNAs not previously described in public databases (Figure 1B).

Table 1. Bioinformatic analysis of FIS2-associated RNAs by RIP-seq. FIS2 was immunoprecipitated from *A. thaliana* siliques and the associated RNA was Illumina sequenced

Library	Raw reads	Trimmed reads	Mapped reads	Recovered reads (%)
Replicate 1	3,574,676	3,549,515	2,694,570	75.2
Replicate 2	3,159,591	3,023,564	2,485,340	81.9

We further classified these FIS2-associated lncRNAs into antisense exonic, intergenic, sense intronic and antisense intronic based on spatial relationships of their loci with protein-coding genes (Figure 1C). Almost half (49%, or 8,239) of the lncRNAs were intergenic and the other 48.9% (8,136) were antisense to exons. A small number (131) were sense intronic lncRNAs and a similar number were transcribed antisense to introns (Figure 1D).

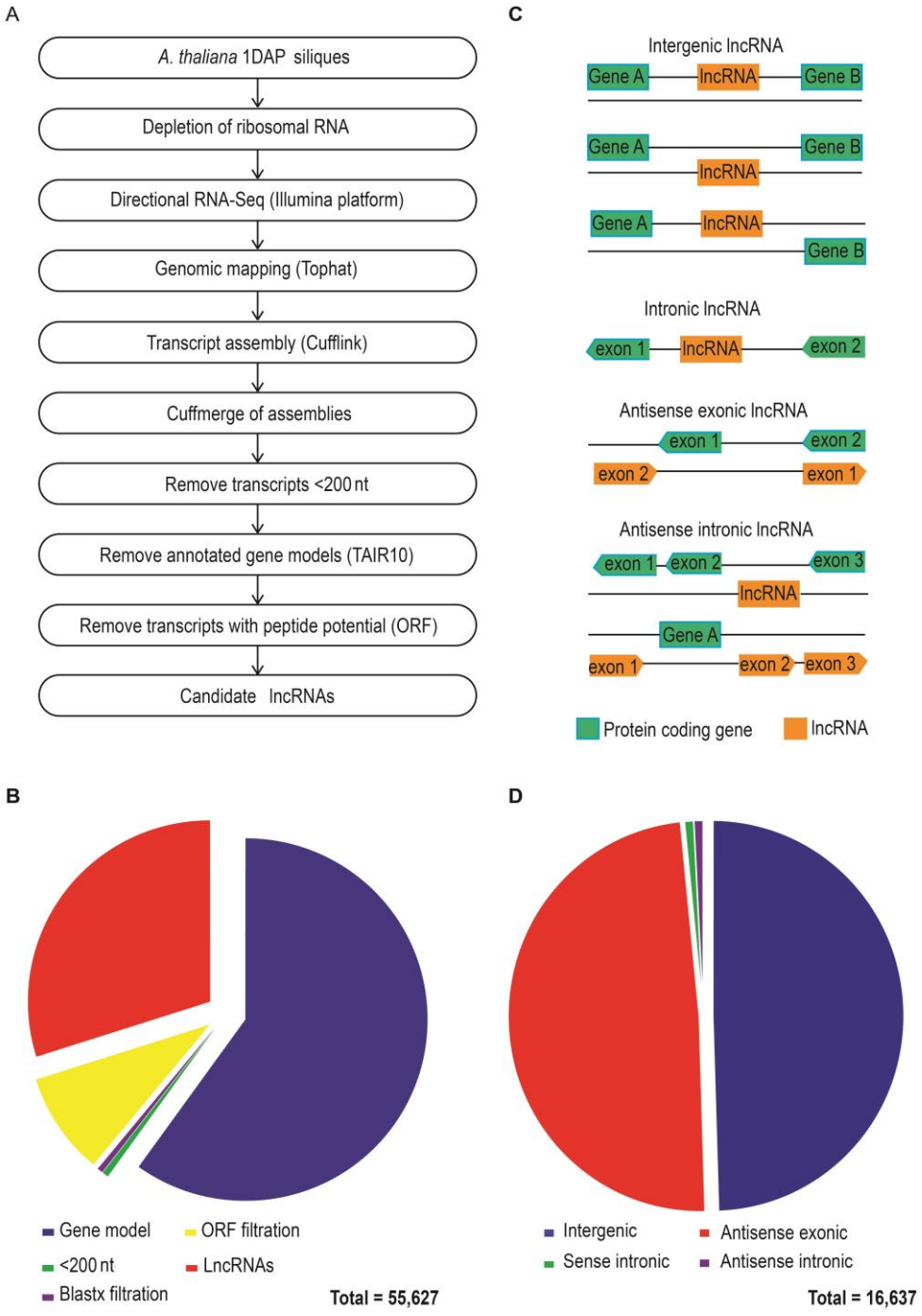


Figure 1. Overview of RIP-seq bioinformatic analysis and characterization of FIS2-associated RNAs. (A) Overview to identify FIS2-associated lncRNAs from 1DAP siliques. (B) *Cuffmerged* transcripts were placed into categories; overlapping with TAIR10 gene models (33,343 transcripts), <200 nt (310 transcripts), homology to *A. thaliana* proteins (Blastx filtration, cut-off E-value ≤0.0001, 273 transcripts), ORF filtration (ORFs >100 amino acids, 5,061 transcripts) and lncRNAs (16,637 transcripts). (C) Schematic classification of FIS2-associated lncRNAs into intergenic, intronic and exonic classes. (D) Classification of FIS2-associated lncRNAs into four

categories: intergenic (8,239 transcripts), sense intronic (131 transcripts), antisense intronic (131 transcripts) and antisense exonic (8,136 transcripts).

Characterization and validation of FIS2-associated lncRNAs

These FIS2-associated lncRNAs have the following characteristics: (1) like protein-coding genes, they are distributed across the five chromosomes with the highest density at the ends of chromosomes (Figure 2B); (2) most have only one or two exons (Figure 2A); (3) they are generally shorter than protein-coding transcripts (Figure 2A); and (4) they have a lower level of expression than protein-coding genes, based on RPKM values (Figure 2A).

We next validated lncRNA–PRC2 interactions by performing RIP-qPCR for five PRC2-associated lncRNAs (*LNC_23526, LNC_23618, LNC_28194, LNC_29066* and *LNC_34938*); *LNC_11274* from outside the PRC2 transcriptome served as a negative control. We found that candidate PRC2-associated lncRNAs are significantly enriched in the HA-tagged FIS2 transgenic lines relative to anti-HA antibody pull-out (Figure 2C & 2D). The negative control showed a band only in input and output lanes, not in the RIP lanes (Figure 2D).

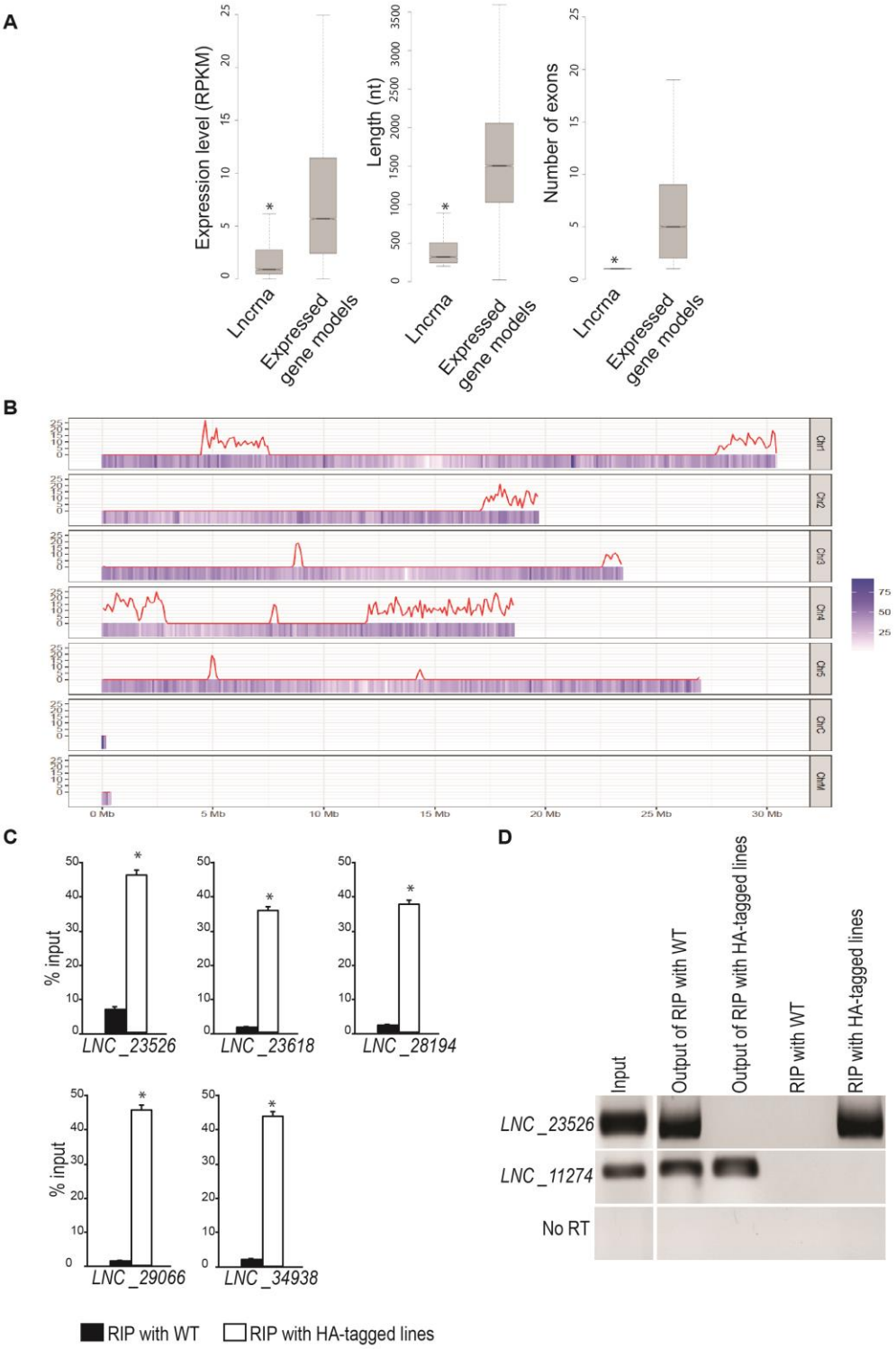


Figure 2. Characterization and confirmation of FIS2-associated lncRNAs. (A) Transcript properties of FIS2-associated lncRNA identified by RIP-seq. (B) Chromosomal distribution of FIS2-associated lncRNAs (red line) and gene models (blue vertical lines). (C) Detection of five FIS2-associated lncRNAs by RT-qPCR after anti-HA IP from wild type (WT) or FIS2:HA transgenic siliques. (D) Confirmation of *LNRNA_11274* not associated with FIS2 (Trung Do et al., in preparation) and FIS2–PRC2-associated *LNCRNA_23526* by RT-qPCR. RT-qPCRs were

performed on two biological and three technical replicates. Error bars indicate ±se of the mean. *P*-values were calculated using Student's t test. Asterisks denote $p < 0.05$.

Conservation analysis of lncRNA–PRC2 interactions

The FIS2–PRC2 complex is specifically expressed in the endosperm and is required for its development [34]. Thus, we first analyzed the conservation of the PRC2 transcriptome by blasting against the genomes of Brassicaceae plants (*B. rapa*, *B. napus* and *B. oleracae*) reported to have FIS2–PRC2 expression. Interestingly, 25, 33 and 21 PRC2-associated lncRNAs were found to share similarities with certain sequences in these respective plant genomes (E-value < 0.001). These numbers mean that around 1% of PRC2-associated lncRNAs have potential conserved homologues. Therefore, the evolutionary conservation of PRC2-associated lncRNAs is low.

In fact, the homologue sequences in other species may or may not encode lncRNAs. Therefore, we proposed to determine whether PRC2-associated lncRNAs are homologous to lncRNAs already identified in these species. Currently, 4,884, 4,403 and 8,594 lncRNAs have been identified in *B. rapa*, *B. napus* and *B. oleracae*, respectively [31]. A total of 16,627 PRC2-associated lncRNAs were blasted against the lncRNAs from those plants, and only two PRC2-associated lncRNAs (*LINC.CUFF.14243.1* and *LINC.CUFF.28728.1*) were found to share similarity—with *CNT0028501* and *CNT0032006*, respectively, in *B. napus* (Table 2). There were no homologues found in *B. rapa* or *B. oleracae*.

Table 2. Identification of conserved lncRNAs in related species

Number of FIS2–PRC2-associated lncRNAs with homologues in other species[#]	Total lncRNAs in other species	Species	Reference for lncRNA identification
0	4,884	*Brassica rapa*	Szcześniak et al., 2016
2	4,403	*Brassica napus*	
0	8,594	*Brassica oleracae*	

[#] The number of FIS2-PRC2-associated lncRNAs with homologues in other species was determined by aligning the 16,627 FIS2–PRC2-associated lncRNAs as queries against the target species' lncRNAs (column 3). The blast E-value cut-off was <0.001.

Functional predictions

We next tested the *cis*-acting functions of PRC2-associated lncRNAs with respect to neighbouring PRC2 target genes by analyzing the RIP-seq data in relation to a hallmark of FIS2–PRC2 activity, H3K27me3. The endosperm H3K27me3 profile was reported by Weinhofer et al. [34], who identified 1,773 H3K27me3 target genes in endosperms. We compared those genes with genes that have PRC2-associated lncRNAs located within 5 kb of their 5'UTR (UTR; untranslated region) or 3'UTR. The results showed that 522 genes overlapped between those datasets (Figure 3A), suggesting proposed functions for lncRNAs in recruiting FIS2–PRC2 complexes to the H3K27me3 target genes. A similar function has been reported for the plant lncRNA *COLDAIR* during vernalisation [5–6, 23].

Further analysis of the overlapping genes showed that 67% were located near the 5' end; the proportion of those at the 3' end was 33% (Figure 3B). This led us to hypothesize a *cis*-acting role of 5’-end PRC2-associated lncRNAs in PRC2 recruitment, adding H3K27me3 marks on neighbouring genes, which are upregulated upon loss of FIS function in 3DAP and 6DAP *fis2* mutant [34].

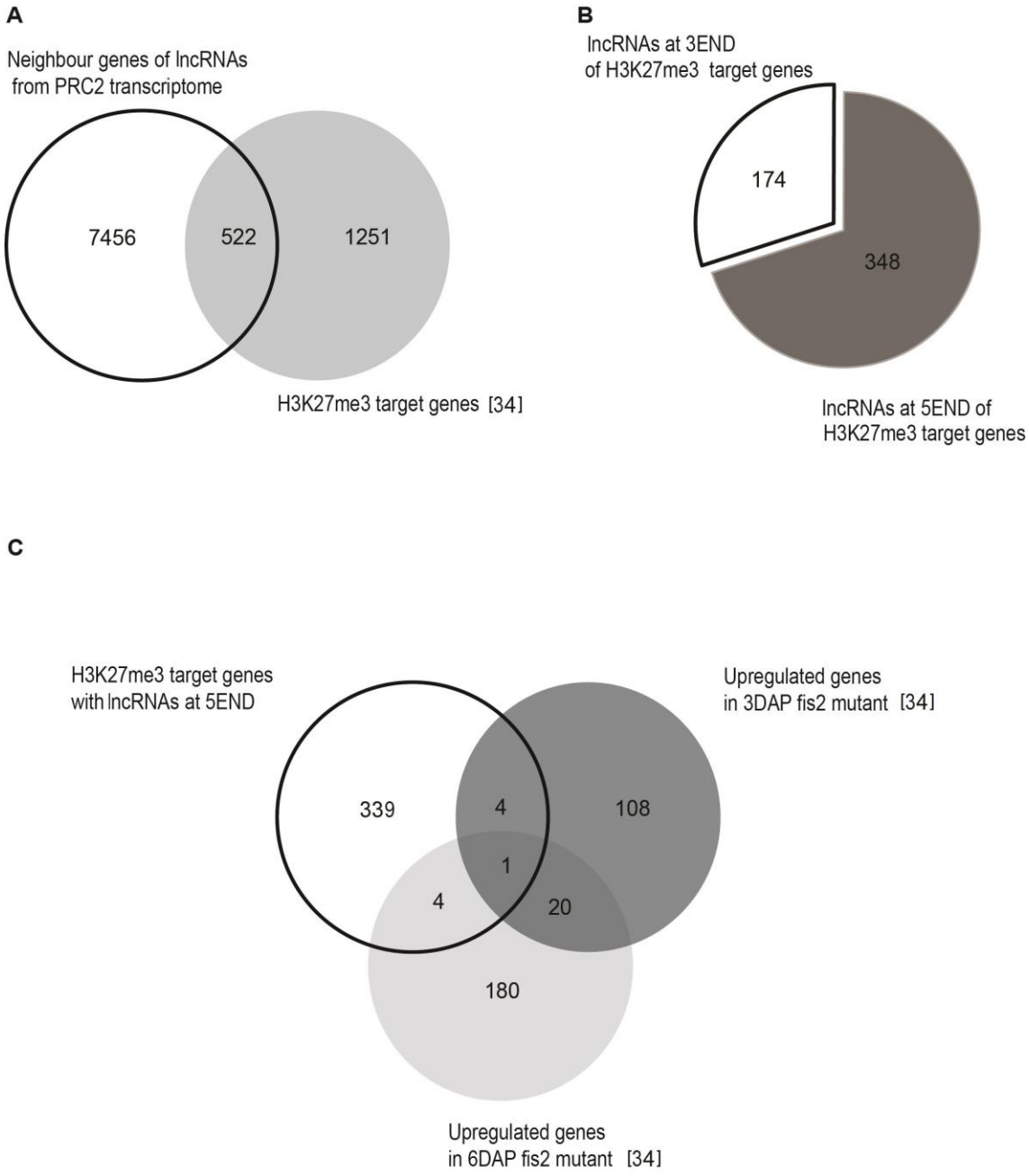


Figure 3. Bioinformatics analysis of PRC2-associated lncRNAs. (A) Overlap between protein-coding genes close to PRC2-associated lncRNAs and H3K27me3 target genes. (B) Proportion of overlap between protein-coding genes close to PRC2-associated lncRNAs and H3K27me3 target genes. (C) Correlation of upregulated H3K27me3 target genes with the protein-coding genes that have lncRNAs located at their 5' end.

To demonstrate this hypothesis, we firstly used publicly available datasets and bioinformatics methods to compare the protein-coding genes that have PRC2-

associated lncRNAs located at their 5' end with upregulated H3K27me3 target genes in 3DAP and 6DAP *fis2* mutant [34]. We showed that 339 genes close to the PRC2-associated lncRNAs did not overlap with data from the *fis2* mutants; only one gene (*AT4G29640*) appeared in all three datasets; and four and four genes overlapped with datasets from the *fis2* mutant at 3DAP and 6DAP, respectively (Figure 3C).

Further, the functions of those nine overlapping genes were also described (Table 3). The data showed that most of the overlapping genes are encoded for enzymes that play important roles during seed development. For example, GDSL-motif lipase/hydrolase family protein preferentially hydrolyse the major component of endosperm cell walls, callose, [35] suggesting that for successful endosperm cellularization the enzymes degrading cell wall need to be silenced.

Table 3. Genes overlapping with RIP-seq lncRNAs, H3K27me3 and upregulated in *fis2* seeds at 3DAP and 6DAP [34]

DAP	Locus	Description
3	*AT1G76500*	DNA-binding family protein
	AT2G25450	2-oxoglutarate-dependent dioxygenase, putative
	AT2G25700	*ARABIDOPSIS* SKP1-LIKE 3
	AT3G59010	Pectinesterase family protein
3 and 6	*AT4G29640*	Cytidine deaminase, putative
6	*AT1G03445*	BRASSINOSTEROID-INSENSITIVE 1 suppressor 1
	AT1G73610	GDSL-motif lipase/hydrolase family protein
	AT1G75900	Family II extracellular lipase 3
	AT1G76290	AMP-dependent synthetase and ligase family protein

Moreover, the relative location of those nine overlapping genes (Table 3) with respect to 5'-end PRC2-associated lncRNAs and H3K27me3 marks in the wild type genome were also confirmed. The results showed that the lncRNAs located at the 5' end of those H3K27me3 target genes and H3K27me3 marks were distributed along the location of H3K27me3 target genes (data not shown). Two

H3K27me3 target genes (*AT1G73610* and *AT1G03445*) and two 5'-end PRC2-associated lncRNAs (*LNC_12840* and LNC_528) are included as representatives in Figure 4.

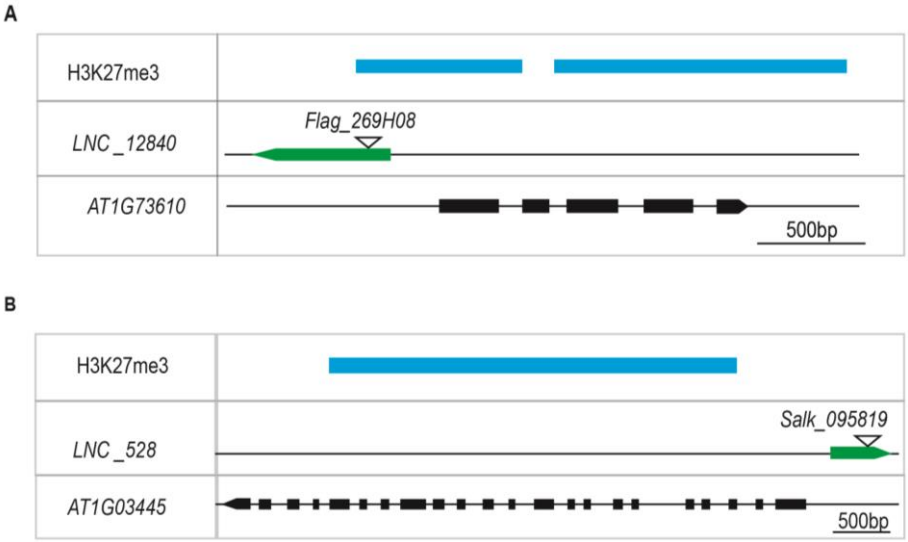


Figure 4. Relative location of genes overlapping with RIP-seq lncRNAs, H3K27me3 and upregulated in fis2 seeds at 3DAP and 6DAP. PRC2-associated lncRNAs (green colour), lnc_12840 (panel A) and lnc_528 (panel B), located at the 5' end of PRC2_target genes (*AT1G73610* (panel A) and *AT1G03445* (panel B)). Histone modification marks, H3K27me3 (blue colour), appear along PRC2_target genes. Mutants for those 5END-lncRNAs were FLAG_269H08 and SALK_095819 (inverted triangle), respectively.

Next, we proposed that mutation in 5'-end of our PRC2-associated lncRNAs would result in the upregulationof target genes (described in Table 3) in mutants. We firstly identified transfer DNA (T-DNA) mutants that could knockdown the expression of those 5'-end lncRNAs and checked the expression of those lncRNAs in these T-DNA mutants by RT-qPCR. As we expected, the results showed that T-DNA insertion led to a transcript reduction in those 5'-end PRC2-associated lncRNAs (Figure 5).

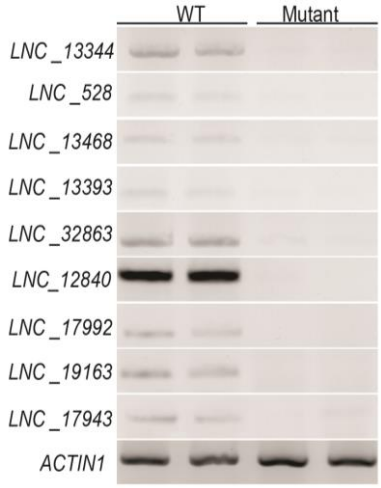


Figure 5. Validation of lncRNAs located at the 5' end of H3K27me3 target genes in mutants for candidate lncRNAs. RT-qPCR analysis of nine lncRNA candidates (lnc_13344, lnc_528, lnc_13468, lnc_13393, lnc_32863, lnc_12840, lnc_17992, lnc_19163, lnc_17943) located at the 5' end of overlapped genes was carried out using material from WT and mutants for those lncRNA candidates. *ACTIN1* was used as an experimental control. RT-qPCRs were performed on two biological and three technical replicates.

Secondly, we measured the mRNA abundance from overlapping genes (see Table 3) in mutants of 5'-end PRC2-associated lncRNAs using RT-qPCR. The results are shown in Table 4.

Table 4. Relative expression of overlapping H3K27me3 target genes in mutant lines for PRC2-associated lncRNAs located at the 5' end of those target genes

DAP	H3K27me3 target genes	5END-lncRNAs	Mutants of 5END-lncRNAs	Relative fold change (±se)
3	*AT1G76500*	*LNC_13468*	*Salk-497543*	0.65±0.315
	AT2G25450	*LNC_17943*	*Flag-205A06*	14.27±0.16
	AT2G25700	*LNC_17992*	*Fag-395F03*	2.78±0.115
	AT3G59010	*LNC_32863*	*Salk-038231*	0.20±0.275
	AT4G29640	*LNC_19163*	*Flag-497A02*	2.10±0.31
6	*AT4G29640*	*LNC_19163*	*Flag-497A02*	1.48±0.215
	AT1G03445	*LNC_528*	*Salk-095819*	38.98±0.07
	AT1G73610	*LNC_12840*	*Flag-269H08*	128.00±0.175
	AT1G75900	*LNC_13344*	*Salk-102768C*	14.27±0.215
	AT1G76290	*LNC_13393*	*Salk-058251*	6.25±0.115

As we expected, the data showed that all overlapping H3K27me3 target genes in 6DAP siliques were upregulated compared with wild type, with fold change values ranging from 1.48 for *AT4G29640* to 128.00 for *AT1G73610*. However, the expression of overlapping H3K27me3 target genes in 3DAP siliques showed a fluctuated change relative to those in the wild type; *AT2G25700*, *AT2G25450* and *AT4G29640* expression were upregulated (fold change 2.78, 14.27 and 2.10, respectively), while *AT1G76500* and *AT3G59010* expression were downregulated (fold change 0.65 and 0.2 respectively).

Collectively, the data indicated that the expression of overlapping H3K27me3 target genes (Table 3) in T-DNA lines is consistent in 6DAP siliques but inconsistent in 3DAP, with their expression in *fis2* mutants. This suggests a possible cis-acting mechanism for predicted lncRNAs in regulating their neighbour genes by binding and guiding the FIS2–PRC2 to the target sites.

Targeting of PRC2 to RNA by short repeats of consecutive guanines

The next question in our study: whether the binding motif is so common that it occurs at a similar frequency in all PRC2-associated lncRNAs. Notably, in humans, Wang et al. [16] reported that PRC2 has a high affinity for folded guanine quadruplex (G-quadruplex) RNA structures and a motif for PRC2-binding RNA composed of short repeats of consecutive guanines. This led us to hypothesize that this motif should be commonly detected in our *A. thaliana* PRC2-associated lncRNAs. To address this, the sequences of 16,637 PRC2-associated lncRNAs were scanned for the presence of putative GQSes. We searched for two or three G-repeats with loop length varying from 1 to 1–3, 1–4 or 1–7 bp (i.e., G2L1, G2L1–2, G2L1–4, G3L1–3 and G3L1–7). The results showed that G2L1–4-type GQSes were most commonly detected, followed by G2L1–2, G2L1, G3L1–7 and G3L1–3 types (Table 5).

Table 5. Number of putative G-quadruplex motifs and motif-containing transcripts identified in 16,627 FIS2–PRC2-associated lncRNAs

	G2L1	**G2L1–2**	**G2L1–4**	**G3L1–3**	**G3L1–7**
No. of putative G-quadruplex motifs	1,896	3,941	11,701	138	545
No. of transcripts for each motif group	1,642	3,225	7,670	136	488

The number of G2L1–4-type GQSes identified was the highest (11,701), while G3L1-3 GQSes were the least common of all GQSes, at only 138 (see Table 5). Notably, more than 90% of the GQSes identified were G2 type; G3 type constituted less than 7% of the total identified GQSes. In addition, the number of transcripts containing G2L1-4 type GQSes was the highest (7,670 transcripts), while the ones containing G3-L1-3 types was the least (136 transcripts).

Overall, our results suggest that G2L1–4 might be a motif in PRC2-binding sites on identified PRC2-associated lncRNAs in *A. thaliana*.

Discussion

Next generation sequencing technologies are very powerful for studying the genome, transcriptome or epigenome of any organism. Plant lncRNAs have been systemically identified in some species [25, 36-40] but most plant transcriptome sequencing data have not been fully explored, leading to the continued lack of understanding of the functions of novel lncRNAs, which may have important roles in a wide range of biological processes [41].

Recently, the significance of lncRNA–protein interactions has been better understood with respect to molecular mechanisms in some biological processes (see Table 2 in Chapter 1). In humans, Khalil et al. [42] used RIP-chip experiments to show that around 20% of the 3,300 lncRNAs expressed in various cell types are bound by SUZ12 or EZH2, two well-known core subunits of PRC2. A similar method was used in mouse embryonic stem cells and approximately 9,000 lncRNAs associated with PRC2 were identified [33]. In this paper, we used a strict computational pipeline and identified 16,637 novel PRC2-associated lncRNAs from 1DAP siliques using a set of *A. thaliana* next generation RIP-seq data. The novel *A. thaliana* PRC2-associated lncRNAs had lower expression levels compared with the mRNAs, which was consistent with previous findings in other species [43-45]. Our conservation analysis showed that among the 16,637 PRC2-associated lncRNAs, only two had homologues in *Brassica napus*. This low level of conservation might be caused by several factors. (1) Current plant lncRNA databases mainly provide nucleotide sequences that are insufficient for conservation of lncRNAs, which may be conserved by structure through species. (2) During lncRNA evolution, each species themselves may have had specific mechanisms to adapt to their habitat; for example, lncRNAs may have short conserved motifs that are not easily identified by BLAST [46] or lncRNAs might encode for small interfering RNAs that are less constrained in other parts of transcripts [43, 46]. (3) There may be factors that affect the formation of a large family with homologous genes: for example, PRC2-associated lncRNAs may interact directly with PRC2 through a conserved secondary structure [16, 23, 47]. In addition, the RT-qPCR results showed that the candidate PRC2-associated lncRNAs are significantly enriched in the HA-tagged FIS2 transgenic lines relative

to anti-HA antibody pull-out. These results provide further evidence that prediction accuracy was sufficient.

Many studies have shown that tissue-specific lncRNAs usually have special functions [36, 48], and the lncRNAs of higher species primarily play the biological role of *cis*-regulation of neighbouring genes [36, 49-51]. In the size range of around 5 kb, we found that 7,988 of 8,239 lncRNA loci had neighbouring protein-coding genes. Therefore, we predicted that the function of these lncRNAs was in PRC2 recruitment, based on the analysis of the H3K27me3 profile of their adjacent coding genes. The results revealed 522 adjacent coding genes that were targets of FIS2–PRC2 complexes but that only around 7 of these might have a *cis*-acting mechanism in regulating the expression of neighbouring genes by PRC2 recruitment.

This small number of H3K27me3 target genes being regulated by interaction of lncRNAs and PRC2 might be because (1) PRC2-associated lncRNAs have multiple functions during developmental stages in which the H3K27me3 marks are associated with active transcription [33, 52]; (2) there are different factors playing roles in PRC2 recruitment to targets, such as DNA-binding transcription factor [53-54] or small RNAs [55-56]; (3) FIS-target genes have stable expression, meaning that polycomb group target genes could be marked by secondary epigenetic modification upon loss of FIS function, or could be suppressed not only by FIS-mediated H3K27me3 but also by other epigenetic modifications that were not removed in the *fis2* mutants; or (4) the PRC2 target genes might be recognised by specific structures that might not be marked by histone modification, such as G-quadruplex structure [16].

Many models have been used to explain lncRNA function whereby lncRNAs have roles as *cis*-acting or *trans*-acting factors to regulate genes at or outside sites where they are transcribed, respectively. In this paper, we suggest a *cis*-acting model for five lncRNAs from PRC2-associated lncRNAs in 6DAP siliques because of the correlation in gene expression between *fis2* mutants and mutants for 5’ end PRC2-associated lncRNAs (Table 4). This model is consistent with another study in which the coordination of lncRNAs and chromatin-modifying PRC2 to target chromatin *FLC* was reported [57]. However, several questions

remain unanswered: (1) we do not know if there are other proteins involved in FIS2–PRC2 recruitment to target chromatin because lncRNAs might act as scaffolds for multiple protein components in this process [58]. For example, the lncRNAs were reported to directly interact with proteins to target chromatin-modifying complexes and guide them to target sites [57-59]. Hence, the identification of the parts of lncRNAs that act as functional motifs is required; (2) lncRNA structure has been reported to play an important role in identifying functions of lncRNAs from plants [57] and animals [60-63]. Therefore, it is important to determine the structure of PRC2-associated lncRNAs that enables them to bind with a chromatin-modifying complex and its target sites.

G-quadruplex structure is one of a variety of three-dimensional structures of DNA inside a cell [64-65]. It is one of the non-canonical four-stranded structures that are made up of multiple Hoogsteen base-paired Guanine-quartets stacked on top of each other [65]. Enrichment of G-quadruplex structures has been found in functional regions of the genome and has been shown to regulate gene expression and translation [66-67]. Recent experiments have established the formation of G-quadruplexes in DNA and RNA in eukaryotic cells [16, 66-67] and plant [68-70]. Interestingly, subsequent results from our sequence and structure analysis showed that G2L1-4 structure might play a significant role in the interaction of lncRNAs with the PRC2 complex in *A. thaliana*. In addition, our result also showed that GQSes with a loop length of 1–3 bp make up the highest proportion, followed by GQSes with loop lengths of 4–5 bp or 6–7 bp. This result is consistent with previous results [70-71] in which G-quadruplexes with shorter loop lengths are more stable than those with longer loop lengths. Further, our results showed that the G2 type was detected more often than the G3 type in the identified lncRNAs, suggesting the PRC2-binding RNA motifs might contain two consecutive guanines. This is consistent with previous results where GQSes in RNA have the potential for transcriptional, translational or mRNA stability regulation [66-70]. Overall, while the actual function of RNA binding by PRC2 is still under investigation by many groups, our finding of low-complexity motifs of short G-tract repeats on PRC-associated lncRNAs provides a means for RNA-mediated regulation of PRC2 in plants.

Lessons from the experimentally functional characterization of some plant lncRNAs indicate the importance of lncRNAs in plant growth and development [36, 48-51]. This has led to the rapid development of genome-wide identification of plant lncRNAs. However, the functional characterization of lncRNAs is lagging far behind and predictions are based on limited methods, including co-expression networks [72], microRNA regulation [73], protein binding [74], epigenetic modification [75] and adjacent gene functions [76]. In this study, we used a method based on epigenetic modification of adjacent genes. Because functional prediction is made via bioinformatics, verification through biological experiments is required to accurately identify the functions of lncRNAs. Their important roles in plant growth and development will be uncovered gradually as biotechnology development continues and more information is published about lncRNAs.

REFERENCES

1. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.
2. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316:1484–1488.
3. Mercer TR, Dinger ME, Mattick JS (2009) Long noncoding RNAs: insights into functions. Nat Rev Genet. 10: 155–159.
4. Wang KC and Chang HY (2011) Molecular mechanisms of long noncoding RNAs. Mol Cell 43 (6): 904-14.
5. Heo JB and Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331: 76–79.
6. Csorba T, Questa JI, Sun Q and Dean C (2014) Antisense COOLAIR mediates the coordinated switching of chromatin states at *FLC* during vernalization. Proc. Natl Acad. Sci. 111: 16160–16165.

7. Davidovich C, Zheng L, Goodrich KJ, Cech TR (2013) Promiscuous RNA binding by Polycomb repressive complex 2. Nat Struct Mol Biol. 20(11): 1250-7.

8. Di Croce L and Helin K (2013) Transcriptional regulation by Polycomb group proteins. Nat Struct Mol Biol. 20(10): 1147-55.

9. Davidovich C, Wang X, Cifuentes-Rojas C, Goodrich KJ, Gooding AR, Lee JT, Cech TR (2015) Toward a Consensus on the Binding Specificity and Promiscuity of PRC2 for RNA. Mol. Cell 57: 552-558.

10. Juan GB & Yukihide T (2015) Cryptic RNA-binding by PRC2 components EZH2 and SUZ12. RNA Biology 12(9): 959-965.

11. Kaneko S, Bonasio R, Saldana-Meyer R, Yoshida T, Son J, Nishino K, Umezawa A, Reinberg D (2014) Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. Mol. Cell 53: 290–300.

12. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311–1323.

13. Kaneko S, Li G, Son J, Xu CF, Margueron R, Neubert TA, Reinberg D (2010) Phosphorylation of the PRC2 component Ezh2 is cell cycle–regulated and up-regulates its binding to ncRNA. Genes Dev. 24: 2615–2620.

14. Kaneko S, Son J, Shen SS, Reinberg D, Bonasio R (2013) PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. Nat. Struct. Mol. Biol. 20: 1258–1264.

15. Klattenhoff CA, Scheuermann JC, Surface LE, Bradley RK, Fields PA, Steinhauser ML, Ding H, Butty VL, Torrey L, Haas S, Abo R, Tabebordbar M, Lee RT, Burge CB, Boyer LA (2013) Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. Cell 152: 570–583.

16. Wang X, Goodrich KJ, Gooding AR, Naeem H, Archer S, Paucek RD, Youmans DT, Cech TR, Davidovich C (2017) Targeting of Polycomb

Repressive Complex 2 to RNA by Short Repeats of Consecutive Guanines. Mol Cell, 65(6):1056-1067.e5.

17. Cifuentes-Rojas C, Alfredo JH, Kavitha S, Jeannie TL (2014) Regulatory Interactions between RNA and Polycomb Repressive Complex 2. Molecular Cell 55 (2): 171-185.

18. Burleigh SH and Harrison MJ (1997) A novel gene whose expression in *Medicago truncatula* roots is suppressed in response to colonization by vesicular-arbuscular mycorrhizal (VAM) fungi and to phosphate nutrition. Plant Mol. Biol. 34(2): 199–208.

19. Liu C, Muchhal US, Raghothama KG (1997) Differential expression of TPS11, a phosphate starvation-induced gene in tomato. Plant Mol. Biol. 33(5): 867–874.

20. Martín AC, Del Pozo JC, Iglesias J, Rubio V, Solano R, De La Peña A, Leyva A, Paz-Ares J (2000) Influence of cytokinins on the expression of phosphate starvation responsive genes in *Arabidopsis*. Plant J. 24(5): 559–567.

21. Wasaki J, Yonetani R, Shinano T, Kai M, Osaki M (2003) Expression of the OsPI1 gene, cloned from rice roots using cDNA microarray, rapidly responds to phosphorus status. New Phytol. 158(2): 239–248.

22. Ding J, Lu Q, Ouyang Y, Mao H, Zhang P, Yao J, et al. (2012) A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. Proc. Natl. Acad. Sci. 109(7): 2654–2659.

23. Swiezewski S, Liu F, Magusin A, Dean C (2009) Cold-induced silencing by long antisense transcripts of an *Arabidopsis* Polycomb target. Nature 462(7274): 799–802.

24. Di C, Yuan J, Wu Y, Li J, Lin H, Hu L, et al. (2014) Characterization of stress-responsive lncRNAs in *Arabidopsis thaliana* by integrating expression, epigenetic and structural features. Plant J. 80(5): 848–861.

25. Wang D, Qu Z, Yang L, Zhang Q, Liu ZH, Do T, Adelson DL, Wang ZY, Searle I, Zhu JK (2017) Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in Plants. Plant J 90:133-146.

26. Xin M, Wang Y, Yao Y, Song N, Hu Z, Qin D, Xie C, Peng H, Ni Z, Sun Q (2011) Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. BMC Plant Biol. 7: 11–61.

27. Zhang H, Chaudhury A, Wu X (2013) Imprinting in plants and its underlying mechanisms. J. Genet. Genomics 40(5): 239-247.

28. David R, Burgess A, Parker, B, Li, J, Pulsford, K, Sibbritt, T, Preiss, T, Searle, I (2017) Transcriptome-wide mapping of RNA 5-methylcytosine in Arabidopsis mRNAs and ncRNAs. The Plant Cell 29(3) 445-460.

29. Burgess AL, David R, Searle IR (2015) Conservation of tRNA and rRNA 5-methylcytosine in the kingdom Plantae. BMC Plant Biol. 15:199.

30. Köster T and Staiger D (2014) RNA-Binding Protein Immunoprecipitation from Whole-Cell Extracts. Methods Mol. Biol. 1062: 679-695.

31. Szcześniak MW, Rosikiewicz W, Makałowska I (2016) CANTATAdb: A Collection of Plant Long Noncoding RNAs. Plant Cell Physiol. 57(1): e8.

32. Huppert, J. L. & Balasubramanian, S (2005) Prevelance of quadruplexes in the human genome. Nucleic Acids Res. 33: 2908–2916.

33. Zhao J, Ohsumi KT, Kung TJ, Ogawa Y, Grau JD, Sarma K, Song JJ, Kingston ER, Borowsky M, Lee TJ (2010) Genome-wide identification of Polycomb-associated RNAs by RIP-seq. Molecular Cell 40: 939–953.

34. Weinhofer I, Hehenberger E, Roszak P, Henning L, Köhler C (2010) H3K27me3 Profiling of the Endosperm Implies Exclusion of Polycomb Group Protein Targeting by DNA Methylation. PLoS Genet. 6(10): e1001152.

35. Minic Z and Jouanin L (2006) Plant glycoside hydrolases involved in cell wall polysaccharide degradation. Plant Physiol. Biochem. 44: 435–449.

36. Song X, Sun L, Luo H, Ma Q, Zhao Y, Pei D (2016) Genome-Wide Identification and Characterization of Long Noncoding RNAs from Mulberry (Morus notabilis) RNA-seq Data. Genes 7: 11.

37. Chen J, Quan M, Zhang D (2015) Genome-wide identification of novel long noncoding RNAs in *Populus tomentosa* tension wood, opposite wood and normal wood xylem by RNA-seq. Planta 241: 125–143.

38. Shuai P, Liang D, Tang S, Zhang Z, Ye C, Su Y, Xia X, Yin W (2014) Genome-wide identification and functional prediction of novel and drought-responsive lincRNAs in *Populus trichocarpa*. J. Exp. Bot. 65(17): 4975-4983.

39. Wu HJ, Wang ZM, Wang M, Wang XJ (2013) Wide spread long noncoding RNAs as endogenous target mimics for microRNAs in plants. Plant Physiol. 161: 1875–1884.

40. Kim ED, Sung S (2012) Long noncoding RNA: Unveiling hidden layer of gene regulatory networks. Trends Plant Sci. 17: 16–21.

41. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 22: 1775–1789.

42. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA 106(28):11667-72

43. Liao Q, Shen J, Liu J, Sun X, Zhao G, Chang Y, Xu L, Li X, Zhao Y, Zheng H (2014) Genome-wide identification and functional annotation of Plasmodium falciparum long noncoding RNAs from RNA-seq data. Parasitol. Res. 113: 1269–1281.

44. Zhang YC, Liao JY, Li ZY, Yu Y, Zhang J, Li Q, Qu L, Shu W, Chen Y (2014) Genome-wide screening and functional analysis identify a large

number of long noncoding RNAs involved in the sexual reproduction of rice. Genome biology. 15(12): 512.

45. Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, Leyva A, Weigel D, García JA, Paz-Ares J (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. Nat. Genet. 39(8):1033–1037.

46. Oosumi T, Gruszewski HA, Blischak LA, Baxter AJ, Wadl PA, Shuman JL, Veilleux RE, Shulaev V (2006) High-efficiency transformation of the diploid strawberry (*Fragaria vesca*) for functional genomics. Planta. 223(6): 1219–1230.

47. Zhang X, Xia J, Lii YE, Barrera-Figueroa BE, Zhou X, Gao S, Lu L, Niu D, Chen Z, Leung C, Wong T, Zhang H, Guo J, Li Y, Liu R, Liang W, Zhu JK, Zhang W, Jin H (2012) Genomewide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function. Genome Biol. 13(3): R20.

48. Grote P, Wittler L, Hendrix D, Koch F, Wahrisch S, Beisaw A, Macura K, Blass G, Kellis M, Werber M, et al. (2013) The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev. Cell 24: 206–214.

49. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida N, Yap CC, Suzuki M, Kawai K, et al. (2005) Antisense transcription in the mammalian transcriptome. Science 309: 1564–1566.

50. Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Soldà G, Simons C (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res. 18: 1433–1445.

51. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS (2008) Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. 105: 716–721.

52. Young MD, Willson TA, Wakefield MJ, Trounson E, Hilton DJ, Blewitt ME, Oshlack A, Majewski IJ (2011) ChIP-seq analysis reveals distinct

H3K27me3 profiles that correlate with transcriptional activity. Nucleic Acids Research, 39(17): 7415–7427.

53. He C, Huang H, Xu L (2013) Mechanisms guiding Polycomb activities during gene silencing in Arabidopsis thaliana. Front Plant Sci. 13(4): 454.

54. Xiao J, Jin R, Yu X, Shen M, Wagner JD, Pai A, Song C, Zhuang M, Klasfeld S, He C, Santos AM, Helliwell C, Pruneda-Paz JL, Kay SA, Lin X, Cui S, Garcia MF, Clarenz O, Goodrich J, Zhang X, Austin RS, Bonasio R, Wagner D (2017) Cis and trans determinants of epigenetic silencing by Polycomb repressive complex 2 in Arabidopsis. Nat Genet. 49(10): 1546-1552.

55. Borges F, Parent JS, van Ex F, Wolff P, Martínez G, Köhler C, Martienssen RA (2018) Transposon-derived small RNAs triggered by miR845 mediate genome dosage response in Arabidopsis. Nat Genet. 50(2): 186-192.

56. Martinez G, Wolff P, Wang Z, Moreno-Romero J, Santos-González J, Conze LL, DeFraia C, Slotkin RK, Köhler C (2018) Paternal easiRNAs regulate parental genome dosage in Arabidopsis. Nat Genet. 50(2): 193-198.

57. Kim D-H, Xi Y, Sung S (2017) Modular function of long noncoding RNA, COLDAIR, in the vernalization response. PloSGenet 13(7): e1006939.

58. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY (2010) Long noncoding RNA as modular scaffold of histone modification complexes. Science 329(5992): 689-693.

59. Jeon Y, Lee JT (2011) YY1 tethers Xist RNA to the inactive X nucleation center. Cell 146(1): 119-133.

60. Postepska-Igielska A, Giwojna A, Gasri-Plotnitsky L, Schmitt N, Dold A, Ginsberg D, Grummt I (2015) LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure. Mol Cell 60(4): 626-636.

61. Kwok CK, Ding Y, Tang Y, Assmann SM, Bevilacqua PC (2013) Determination of in vivo RNA structure in low-abundance transcripts. Nat Commun 4: 2971.

62. Ding Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505(7485): 696-700.

63. Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. Nature 505(7485): 701-705.

64. Phan AT, Kuryavyi V and Patel DJ (2006) DNA architecture: from G to Z. Curr. Opin. Struct. Biol. 16:288–298.

65. Bochman ML, Paeschke K and Zakian VA (2012) DNA secondary structures: stability and function of G-quadruplex structures. Nat. Rev. Genet. 13:770–780.

66. Huppert JL & Balasubramanian S (2005) Prevelance of quadruplexes in the human genome. Nucleic Acids Res. 33, 2908–2916.

67. Murat P and Balasubramanian S (2014) Existence and consequences of G-quadruplex structures in DNA. Curr. Opin. Genet. Dev. 25:22-29.

68. Garg R, Aggarwal J, Thakkar B (2016) Genome-wide discovery of G-quadruplex forming sequences and their functional relevance in plants 6:28211.

69. Mullen MA, Olson KJ, Dallaire P, Major F, Assmann SM, Bevilacqua PC (2010) RNA G-Quadruplexes in the model plant species Arabidopsis thaliana: prevalence and possible functional roles. Nucleic Acids Res. 38:8149–8163.

70. Takahashi H, Nakagawa A, Kojima S, Takahashi A, Cha BY, Woo JT, Nagai K, Machida Y, Machida C (2012) Discovery of novel rules for G-quadruplex-forming sequences in plants by using bioinformatics methods. J. Biosci. Bioeng. 114:570.

71. Bugaut, A. & Balasubramanian, S (2008) A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. Biochemistry 47, 689–697.

72. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk Q, Carey BW, Cassady JP (2009) Chromatin signature reveals over a

thousand highly conserved large noncoding RNAs in mammals. Nature 458: 223–227.

73. Keniry A, Oxley D, Monnier P, Kyba M, Dandolo L, Smits G, Reik W (2012) The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. Nat. Biol.14: 659–665.

74. Yang JH, L JH, Jiang S, Zhou H, Qu LH (2013) ChIPBase: A database for decoding the transcriptional regulation of long noncoding RNA and microRNA genes from ChIP-Seq data. Nucleic Acids Rese. 41: D177–D187.

75. Sati S, Ghosh S, Jain V, Scaria V, Sengupta S (2012) Genome-wide analysis reveals distinct patterns of epigenetic features in long noncoding RNA loci. Nucleic Acids Res. 40: 10018–10031.

76. Boerner S and McGinnis KM (2012) Computational identification and functional predictions of long noncoding RNA in Zea mays. PLoS ONE 7(8): e43047.

Chapter 6: Maternal Control of Seed Size by a Long noncoding RNA in *Arabidopsis thaliana*

Trung Do[1], Zhipeng Qu[1], Ashley Jones[1], Rakesh David[2], Dave Adelson[1], and Iain Searle[1,2*]

[1] Department of Molecular and Cellular Biology, School of Biological Sciences, The University of Adelaide, Adelaide, South Australia, 5005, Australia.

[2] School of Agriculture Food and Wine, The University of Adelaide, Adelaide, South Australia 5005, Australia.

[*] Corresponding author: iain.searle@adelaide.edu.au

In preparation for Plant Cell

Statement of Authorship

Title of paper	Maternal control of seed size by a long non-coding RNA in *Arabidopsis thaliana*
Publication Status	In preparation
Publication details	To be submitted to The Plant Cell

Author Contributions

By signing the Statement of Authorship, each author certifies that:

iv. The candidate's stated contribution to the publication is accurate (as detailed below)
v. Permission is granted for the candidate in include the publication in the thesis and
vi. The sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Candidate	Trung Q. Do		
Contribution to paper	Co-designer of experiment. Analysed the Fluidigm Access Array data. Generated the mutants for lncRNA_1246 and AT3G12940. Wrote the manuscript and composed figures.		
Overall percentage (%)	70%		
Signature		Date	29/05/18

Name of Co-Author	Zhipeng Qu		
Contribution to paper	Bioinformatic analysis of RNA-seq data. Wrote the Materials and Methods section on bioinformatics analysis. Overall contribution 15%.		
Signature		Date	30/05/2018

Name of Co-Author	Ashley Jones		
Contribution to paper	Generated and partially characterized MPC-NTF transgenic line. Overall contribution 2%.		
Signature		Date	6/6/18

Name of Co-Author	Rakesh David		
Contribution to paper	Contributed to method development. Overall contribution 2%.		
Signature		Date	1/6/18

Name of Co-Author	David L. Adelson		
Contribution to paper	Supervised development of bioinformatics analysis. Overall contribution 2%.		
Signature		Date	1/6/18

Name of Co-Author	Iain R. Searle		
Contribution to paper	Co-designer of all experiments. Supervised development of work and edited the manuscript. Overall contribution 9%.		
Signature		Date	29/5/18

Abstract

Background: Human nutrition is mainly derived from cereal grains or pulses. Therefore, studying the genetic mechanisms that control seed size is important as it may allow us to modify the regulators to increase final seed size and nutrient intake. While many protein-coding genes have been identified to have an important function in seed development, the roles of non-coding RNAs are largely unexplored. A few long non-coding RNAs (lncRNAs) have been shown to play important regulatory roles in post-transcriptional and transcriptional regulation, but most have no clear functional role. No lncRNAs have yet been demonstrated to play a functional role in seed development in plants.

Results: We purified *Arabidopsis thaliana* endosperm nuclei 24 h after pollination using the INTACT protocol, deep-sequenced the RNA and identified 31,608 transcripts. Of these transcripts, 615 were annotated as lncRNAs, of which more than 80% contained a single exon and were shorter than 400 nucleotides. We determined the tissue expression pattern of five of these lncRNAs using Fluidigm arrays and demonstrated they were tissue-specifically expressed. Next, we knocked down one of these—the nuclear-specific, antisense *LNCRNA_1246*—using strand-specific artificial microRNAs and found that mutant plants had smaller cells and organs, including seeds. Finally, using reciprocal crosses, we demonstrated that *LNCRNA_1246* acts maternally to control seed size by possibly regulating outer integument cell size.

Conclusions: Our data are the first to demonstrate that a lncRNA can control cell and organ size in plants.

Keywords: *Arabidopsis thaliana*, cell size, integument, long non-coding RNA, maternal effect, seed size.

Introduction

Humankind is facing increasing food insecurity because of overpopulation, climate change and the increasing demand for fertile land to raise biofuel crops. Therefore, a major challenge for the 21st century is to successfully apply our current knowledge and new approaches to sustainably increase crop yields. As

60% of the calories that humans consume come from cereal grains and a significant proportion of amino acids are derived from pulse grains, understanding the mechanisms that regulate endosperm development and final seed size are of fundamental importance in addressing food security. One useful model plant to study seed development is *Arabidopsis thaliana*, as plants are easily cultivated, there are extensive genetic and community resources and plants are easily transformed. To date, only a handful of genes have been reported to have direct involvement in determining *Arabidopsis* seed size [1–8].

Flowering plant seeds are derived from two fertilization events that often occur deep within the female gametophyte: one sperm cell fertilizes the egg cell to form the embryo, and the second sperm cell fertilizes the central cell to produce the endosperm. The endosperm usually surrounds the embryo and in turn is enclosed within the ovule integument. Therefore, producing a mature three-dimensional seed requires coordinated cell division, cell expansion and inter-cellular communication during seed development [9].

Endosperm and embryo development is also under epigenetic control [10-12]. Mutations in *FERTILIZATION-INDEPENDENT SEED 2 - POLYCOMB RECESSIVE COMPLEX 2 (FIS2-PRC2)* genes result in autonomous and excessive endosperm development and seed abortion [11-12]. In *fis* mutant seeds, embryos arrest at the heart stage and endosperm cells do not cellularize, resulting in additional cell proliferation during late development than in wild type [11-12]. The FIS–PcG complex in *Arabidopsis* consists of four proteins—MEDEA (MEA), FERTILIZATION-INDEPENDENT ENDOSPERM (FIE), FERTILIZATION-INDEPENDENT SEED2 (FIS2), and MULTI-COPY OF IRA1 (MSI1)—that control endosperm development through depositing repressive H3K27me3 histone modifications at imprinted loci [10]. Regulatory non-coding RNAs associated with FIS–PcG were explored by Trung Do et al. (in prep); however their function is still to be fully elucidated.

To identify non-coding RNAs transcribed early in endosperm development, we constructed a transgenic *A. thaliana* line expressing the INTACT fusion protein under the control of the *MATERNALLY EXPRESSED PAB C-TERMINAL (MPC)* promoter in developing endosperm cells. After purification of endosperm nuclei

24 h after pollination and Illumina sequencing, we identified 615 novel lncRNAs. Knockdown of one lncRNA, *LNCRNA_1246*, resulted in decreased seed size by reducing the outer integument cell size. Via reciprocal crosses we also demonstrated that *LNCRNA_1246* acts maternally to control seed size. Further, we showed that all tested cells and organs were smaller than wild type.

Materials and methods

Plant materials

A. thaliana (Columbia-0 accession) wild type and transgenic lines were grown in Phoenix Biosystems growth under metal halide lights as previously described [13]. For plate experiments, seeds were surface sterilized for 12 h using chlorine gas, plated on ½ MS medium supplemented with 1% sucrose and sealed as previously described [14]. All plants were grown under long-day photoperiod conditions of 16 h light and 8 h darkness at an ambient temperature of 21^{o}C.

Homozygous MPC promoter::nuclear targeted fusion (pMPC::NTF) transgenic lines were constructed by Ashley Jones (The Australian National University) and transformed into Columbia wild type. Transgenic plants were selected on ½ MS media supplemented with 15 μg ml^{-1} hygromycin B. NTF transcript abundance was assessed in at least five independent T_1 plants using quantitative reverse transcription PCR (RT-qPCR) and two lines with NTF mRNA abundance similar to wild type transcript levels were carried through to the homozygous T_3 generation for molecular analysis. Siliques from pMPC::NTF or wild type were hand pollinated and harvested 1 day later for nuclei purification and RNA isolation experiments.

RNA isolation and library construction for RNA-seq

Siliques from two biological replicates of either wild type or pMPC::NTF 1 DAP (1 day after pollination) plants were snap frozen in liquid nitrogen and used immediately or stored at –80^{o}C. One gram of siliques was ground to a fine powder using a mortar and pestle and their nuclei purified as previously described [15-16].

To construct RNA-seq libraries, whole cell lysates were prepared from formaldehyde-fixed siliques, treated with 400 U/ml DNAse I (NEB) and 20 U/ml RNaseOUT™ (ThermoFisher Scientific) and incubated with streptavidin beads (ThermoFisher Scientific) for 30 min. Total RNA was extracted using TRIzol (ThermoFisher Scientific) and rRNA was removed using the Ribo-Zero rRNA Removal Kit (Plant Seed/Root) (Ilumina®, product code MRZSR116). Strand-specific RNA-seq libraries were constructed using the NEBNext Ultra Directional RNA library prep kit (NEB) and sequenced using an Illumina HiSeq™ 2500 in paired-end (PE100) mode.

Transcriptome detection by RNA-Seq

Adaptor and low-quality sequences of raw reads were removed using trim_galore with the parameter -- stringency 6, and trimmed reads were then aligned against the *A. thaliana* TAIR10 genome assembly using *TopHat2* with the following parameters: -N 5 -- read-edit- dist 5. Aligned reads from all samples were merged with SAMtools and assembled using *Cufflinks* with the following parameter: -- library-type fr-firststrand -u. Assembled transcripts were filtered through our lncRNA identification pipeline as previously described [17]. Transcripts shorter than 200 nucleotides (nts) were removed and genomic coordinates of long transcripts were checked against reference genes of TAIR10 and classified into either gene transcripts; intergenic transcripts; intronic transcripts; or antisense transcripts. The latter three classes of transcript were selected to filter unannotated protein-coding potential transcripts using the following two steps: 1) sequence similarity search against the Swiss-Prot protein database; and then 2) predicted open reading frame(s) (ORF).

Quantitative reverse transcription PCR

Validation of expression was performed as described [14]. Nuclei purification was followed by quantitative, strand-specific RT-PCR (RT-qPCR). First-strand cDNA synthesis was carried out with approximately 300 ng RNA using the SuperScript™ III (ThermoFisher Scientific). All primers used in this study are listed in Table S1. RT-qPCR was performed in quadruplicate using the SYBR Green Mastermix (Roche Applied Science) on a Fluidigm BioMark™ HD

(Fluidigm®) according to the manufacturer's instructions. Sample cycle threshold (Ct) values were determined and standardized relative to the input, and the $2^{-\Delta CT}$ method was used to calculate the relative change in gene expression based on the RT-qPCR data.

Plasmid construction and generation of transgenic plants

The artificial microRNA (amiRNA) sequences used for the strand-specific knockdown of lncRNAs were synthesized by Integrated DNA Technologies and cloned into Gateway entry vector pENTR/D (ThermoFisher Scientific) according to the manufacturer's instructions. Inserts were Sanger sequenced and then cloned into the destination vector pLEELA using the Gateway cloning system [18] by following the manufacturer's instructions. The resulting constructs were driven by the strong CaMV35S promoter. The constructs were transformed into *Arabidopsis* wild type Col-0 plants by the *Agrobacterium tumefaciens*-mediated floral dip method [19]. Transgenic plants were selected on soil by spraying with BASTA (120 mg/L). In transgenic lines, amiRNA transcript abundance was assessed in at least five independent T_1 plants using RT-qPCR, and two lines showing the highest amiRNA transcript levels were carried through to the homozygous T_3 generation for phenotypic and molecular analysis.

Cross-pollination experiments

For reciprocal crosses between individual *Arabidopsis* plants, the anthers were emasculated before bud opening and covered with ClingWrap (Glad® Foil) for 36 h until the stigma was mature, to generate the female parent. Mature pollen from the pollen donor parent was applied to receptive stigmas under a dissecting microscope. After pollination, female parents were returned to the plant growth chambers.

Mature seed weight measurements

The mature seeds were separated from the dry siliques using a sieve and the seeds were stored for 2 weeks in a sealed container with silica before weighing. Three batches of 100 seeds were weigh using an electronic scale (AG204 DeltaRange®, product model AG204DR).

Confocal laser-scanning microscopy

To measure outer integument cell size of the ovules, the ovules before and 1 DAP were fixed with 4% glutaraldehyde and cleared with benzyl benzoate: benzyl alcohol (2:1 v/v). Images were taken using an Olympus FV3000 confocal laser-scanning microscope. Excitation wavelengths were 488 nm for green fluorescent protein (GFP) and a collection range of 488–700 nm was used. Cell area measurements were performed using imageJ. The measured area were converted into cell volume using the Microsoft Excel 2016 software.

Results

Transcriptome-wide identification of lncRNAs from *Arabidopsis* endosperm nuclei

In both the plant and animal kingdoms, it is becoming clearer that the non-coding RNA regulatory network is important for cell identity and developmental processes [20]. However, the identity and functional roles of lncRNAs in plant endosperm development are largely unknown. To address this important question, we applied a purification protocol for nuclei from early stages of endosperm development of *A. thaliana* and sequenced the RNAs.

To purify endosperm nuclei, we used the INTACT system [15-16]. In *Arabidopsis*, MPC expression is restricted to the female gametophyte and developing endosperm tissue [21]; using the MPC promoter to drive the NTF protein led to labelling of endosperm nuclei. The labelled nuclei were subsequently biotinylated *in vivo* and purified from the total nuclear pool by virtue of the high-affinity interaction between biotin and streptavidin [12-13]. We produced a single insert, pMPC::NTF transgenic line and applied a stringent purification protocol using streptavidin-coated magnetic beads as described previously [12-13]. Briefly, after applying our stringent purification protocol to wild type siliques, no 4′,6-diamidino-2-phenylindole (DAPI)-stained nuclei were detected using fluorescence microscopy (data not shown). To demonstrate that no substantial contaminating RNA was associated with the beads, we constructed an Illumina library; however, no inserts were detected between the adapters after Bioanalyzer analysis. In

contrast, after applying the purification protocol to pMPC::NTF siliques, many DAPI-stained nuclei were detected among the streptavidin-coated beads. Therefore, we concluded that using our purification protocol and stringent washes, only NTF-labelled endosperm nuclei would be purified, and associated RNAs sequenced.

We harvested biological replicates of silique tissue from transgenic pMPC::NTF plants (Figure 1A), purified the endosperm nuclei, isolated and purified the associated RNAs, constructed libraries and Illumina sequenced the libraries. The sequenced libraries were analyzed by our bioinformatic pipeline to annotate the lncRNAs (Figure 1B). Briefly, sequence reads were aligned using *TopHat2* to preserve junction reads (Table 1) and then *Cufflinks* was used to assemble uniquely mapped reads into known and novel transcripts. These transcripts were combined using *Cuffmerge* and then compared with the reference annotation using *Cuffcompare*. Based on the huge availability of deep-sequencing datasets, the merged transcripts were compared with known protein-coding genes and lncRNAs in the public databases to obtain a minimum number of novel, false positive lncRNA transcripts (Figure 1C). We identified 35,097 reliably expressed transcripts and among them, 34,531 were longer than 200 nt. Of these 34,531 transcripts, 615 were identified as lncRNAs not previously described in public databases (Figure 1C).

Table 1. Bioinformatic analysis of endosperm-associated RNAs by RNA-seq. Endosperm nuclei were purified from *A. thaliana* siliques and the associated RNA was Illumina sequenced

pMPC::NTF samples	Raw reads	Cleaned reads	Mean length (bp)	Mapped (%)
Replicate 1	2,421,561	2,300,671	97.02	95.12
Replicate 2	2,407,807	2,274,067	96.72	94.58

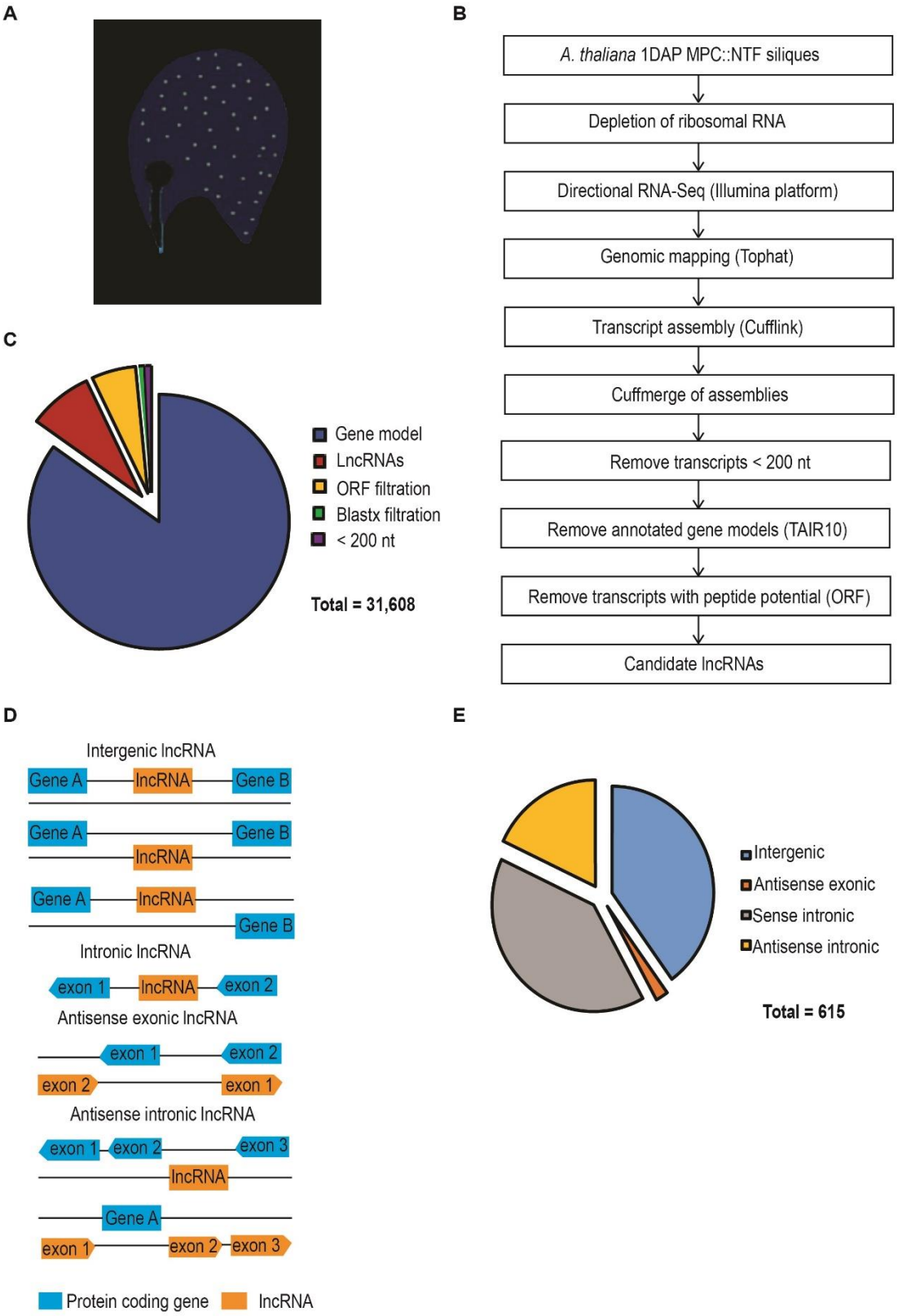


Figure 1. Overview of RNA-seq bioinformatics analysis and characterization. A) Cartoon of the expression of nuclear targeted fusion (NTF) protein driven by the *MATERNALLY EXPRESSED PAB C-TERMINAL (MPC)* promoter in endosperm nuclei of *Arabidopsis thaliana* seeds (green

dots). B) Overview to identify endosperm-associated long non-coding RNAs (lncRNAs) from early endosperm tissue. C) *Cuffmerged* transcripts were characterized into categories: overlapping with TAIR10 gene models (33,567 transcripts); <200 nucleotide (nt) (566 transcripts); homology to *A. thaliana* proteins (Blastx filtration, cut-off E-value ≤ 0.0001, 28 transcripts); open reading frame (ORF) filtration (ORFs > 100 amino acids, 321 transcripts); and lncRNAs (615 transcripts). D) Diagram for the classification of lncRNAs from 1 day after pollination (1DAP) siliques. E) Classification of endosperm-associated lncRNAs into four categories: intergenic (258 transcripts); sense intronic (257 transcripts); antisense intronic (1 transcripts); and antisense exonic (99 transcripts).

Next we asked whether there was an association between the lncRNAs and the biological function of nearby genes. We observed no clear relationship as determined by a Gene Expression Omnibus biological function term analysis (data not shown). We next classified the lncRNAs based on their genomic position with respect to protein-coding genes (Figure 1D). Exonic antisense and intergenic lncRNAs are the two largest classes, with 257 and 258 transcripts, respectively (Figure 1E).

Characterization of the identified endosperm-associated lncRNAs

Protein-coding genes in the *A. thaliana* genome are distributed across the five chromosomes, with lower abundance around the centromeres (Figure 2A). Similar to the protein-coding gene distribution, the lncRNAs from our dataset were also distributed across all chromosomes (Figure 2A). In contrast to protein-coding genes, the lncRNA loci had lower expression levels, based on the RPKM (Reads Per Kilobase Million) value calculated by *Cufflink* (Figure 2B) and their transcript length was shorter (Figure 2C). The length distribution of most lncRNAs is in the range of 200–500 bp, whereas the transcript length for protein-coding genes is mostly above 800 bp. Additionally, most lncRNAs, 85%, have only one or two exons while only 34% of protein-coding genes have fewer than two exons.

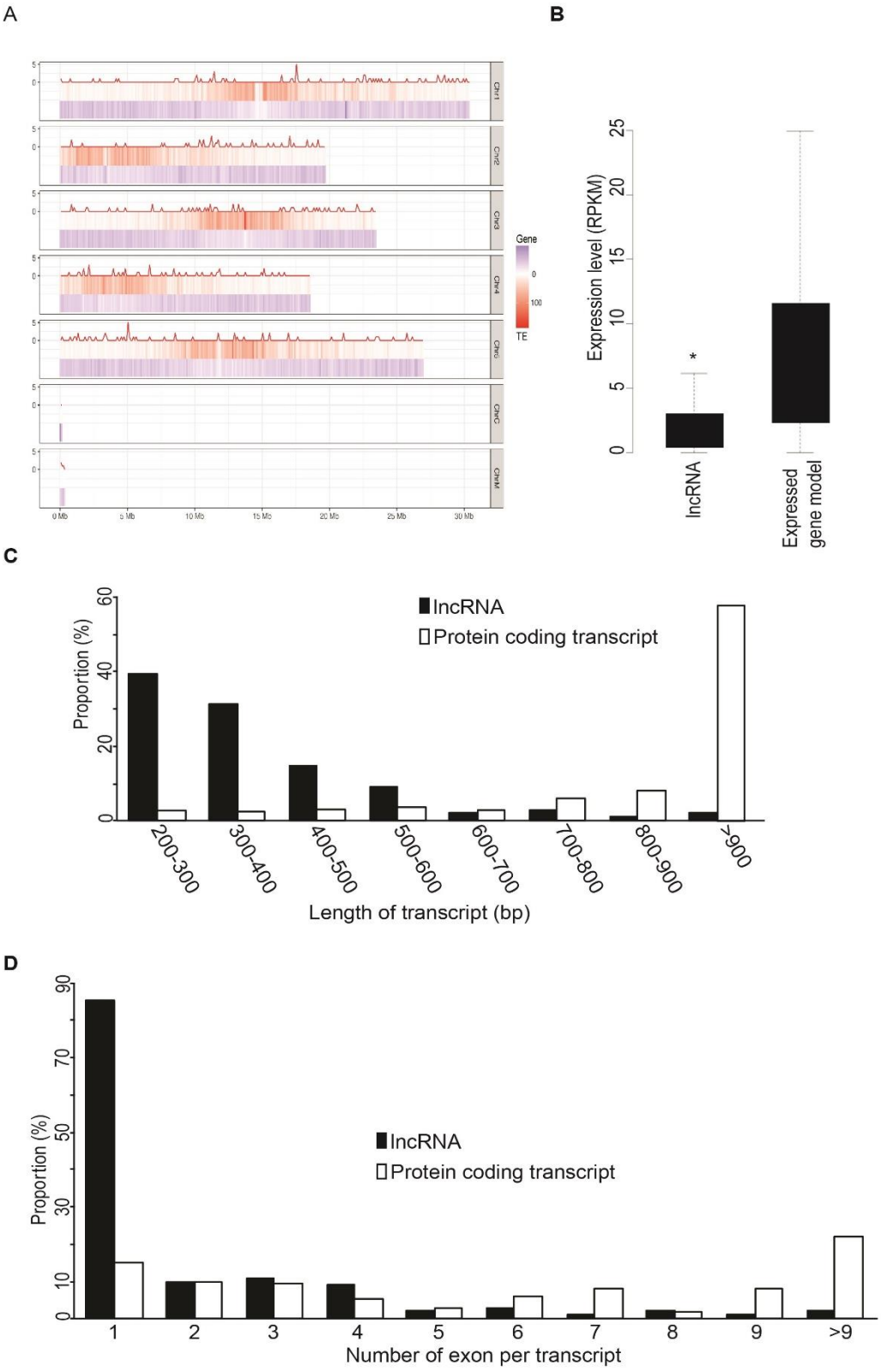


Figure 2. Characterization of endosperm-associated long non-coding RNAs (lncRNAs). A) Chromosomal distribution of endosperm-associated lncRNAs (red line), transposable elements (TE, red vertical lines) and gene models (blue vertical lines). B) Boxplot showing transcript

abundance of lncRNAs and gene models in *Arabidopsis* endosperm (*$p \leq 0.0001$, Spearman's correlation). C) and D) Transcript properties of endosperm-associated lncRNAs identified by RNA-seq.

Validation of endosperm-associated lncRNAs

To validate the RNA-seq results, RT-qPCR was performed for five endosperm-associated lncRNAs identified in the present study (*LINC.TCONS_2215*, *EXONAS.TCONS_244, INTRONAS.TCONS_1171, EXONAS.TCONS_1177* and *INTRONAS.TCONS_682*) and another five lncRNAs (*INTRONAS.TCONS_120, LINC.TCONS_719, INTRONAS.TCONS_976, INTRONAS.TCONS_2762, INTRONAS.TCONS_2182*) from other datasets (Trung Do et al., in prep). The results showed that all of the endosperm-associated lncRNAs have a similar expression pattern, with higher abundance in siliques than in root or floral tissue (Figure 3A). The non-endosperm-associated lncRNAs were either not detected or were in low abundance in silique tissue, but were in higher abundance in root and floral tissue (Figure 3A).

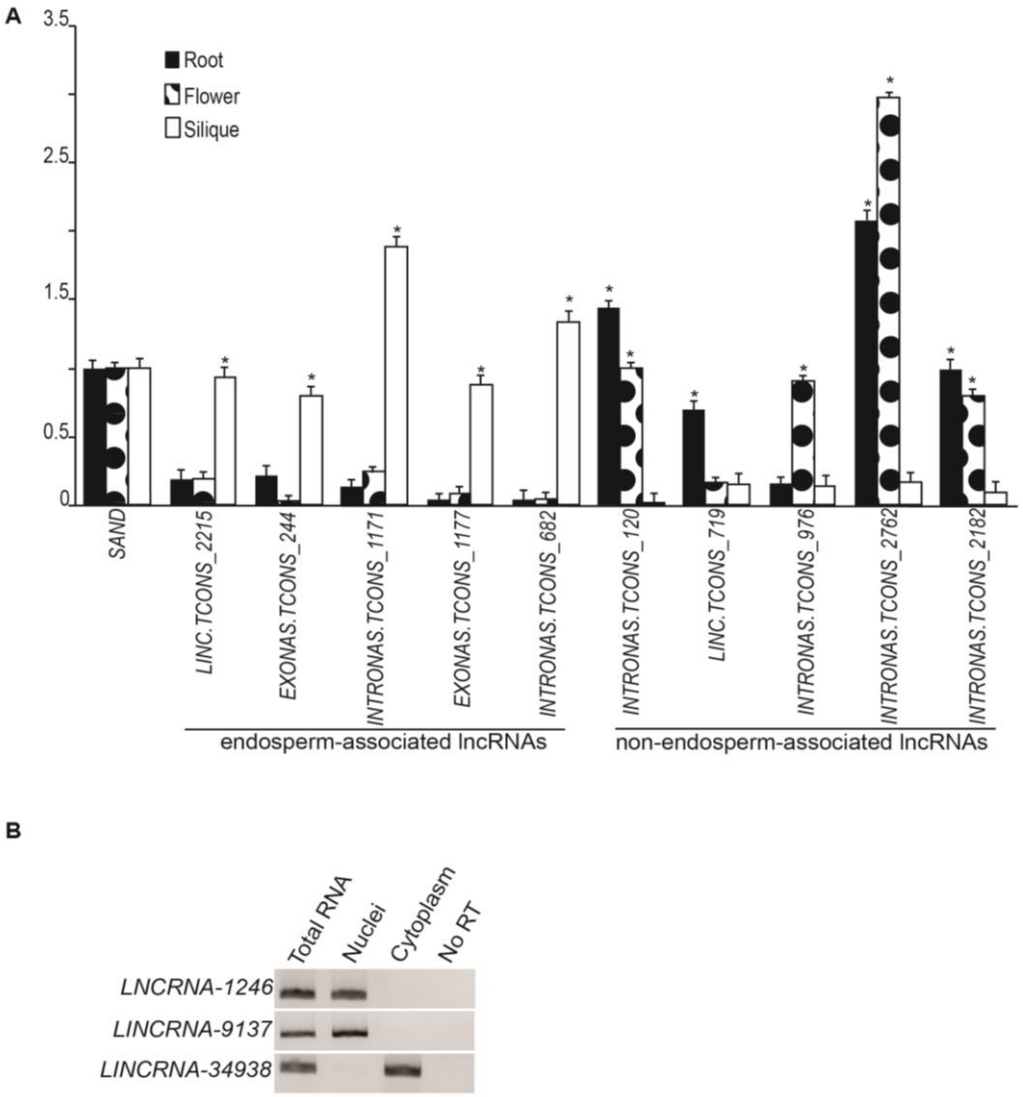


Figure 3. Tissue expression patterns and sub-cellular localization of long non-coding RNAs (lncRNAs). A) Confirmation of five endosperm-associated lncRNAs and five lncRNAs identified in other datasets by RT-qPCR. The RT-qPCR data for all transcripts were normalised to the housekeeping gene SAND. RT-qPCRs were performed on two biological and three technical replicates. Error bars indicate ±SE of the mean. *p*-values were calculated with a Student's *t*-test. An asterisk denotes $p < 0.05$. B) PCR amplification of lncRNAs from RNA purified from either nuclei, cytoplasm or total RNA. *LNCRNA_1246* was amplified from total and nuclear RNA but not from cytoplasmic RNA. *LNCRNA_9137* was previously identified as a nuclear-specific lncRNA [17] and *LINCRNA_34938* as a cytoplasmic-specific lncRNA (Trung Do et al., in prep).

Next, we asked whether one endosperm-associated lncRNA, *LNCRNA_1246*, was preferentially enriched in either the cytoplasm or nuclear sub-cellular compartments. To do this, we isolated nuclei and cytoplasmic fractions, purified the RNA and converted the RNA to cDNA. First, to test the purity of our nuclear

and cytoplasmic fractions, we measured the abundance of the nuclear-specific *LNCRNA_9137* [17] and cytoplasmic-enriched *LINCRNA_34938* (Trung Do et al., in prep). We did not detect any lncRNA_9137 transcripts in the cytoplasm, nor any *lncRNA_34938* transcripts in the nucleus and concluded our preparations had no detectable contamination (Figure 3B). Next, we measured the abundance of *LNCRNA_1246* in the nuclear and cytoplasmic fractions and detected it only in the nuclear fraction (Figure 3B).

Arabidopsis thaliana LNCRNA_1246 mutants have smaller seeds than wild type

We focused our functional analysis on the nuclear-specific *LNCRNA_1246*. We generated transgenic plants containing two strand-specific amiRNAs, *lnc1246-1* and *lnc1246-2*, and isolated a homozygous T-DNA insertion, *Salk_207384* (called *lnc1246-3* in Figures) (Figure 4A). All three mutants had an approximate 30% reduction in seed weight compared with the wild type (Figure 4D & 4F). Next, we measured the abundance of *LNCRNA_1246* in RNA isolated from seeds of the three mutants; as expected we detected no *LNCRNA_1246* transcripts (Figure 4B).

As *LNCRNA_1246* is an antisense transcript of a protein-coding gene, *AT3G12940*, we further questioned whether the smaller seeds in the mutants resulted from mutation of *LNCRNA_1246* or *AT3G12940*. To address this, we strand-specifically knocked down *AT3G12940* using an amiRNA, named here *At3g12940-1* (Figure 4A). As expected, mutant seeds of *At3g12940-1* were the same weight and size as wild type (Figure 4E) and we could not detect *AT3G12940* mRNA; however we could detect the antisense transcript *LNCRNA_1246* in the mutants. In summary, only *lnc1246* mutants have reduced seed weight; *At3g12940* mutants have the same seed weight and size as wild type (Figure 4C).

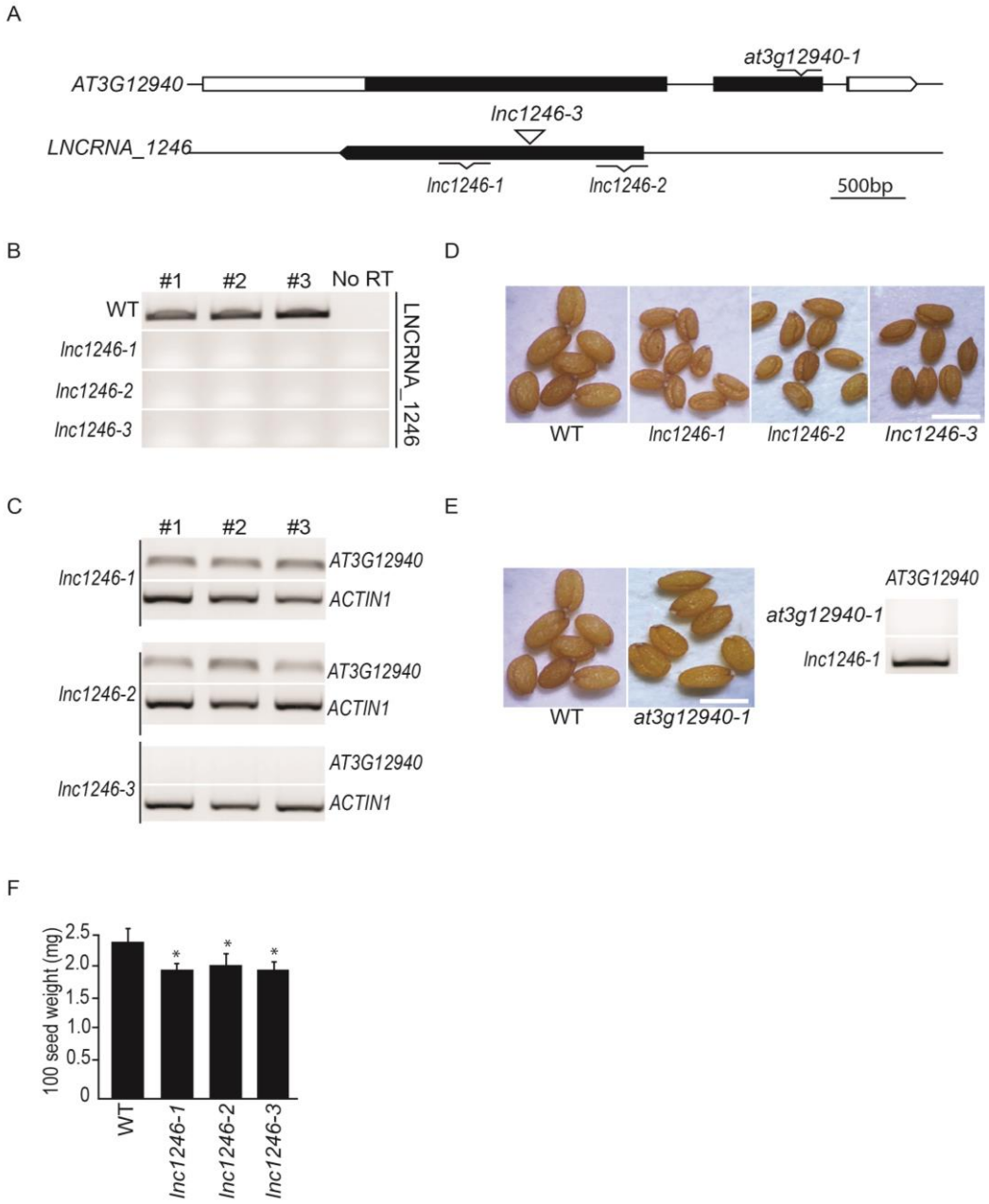


Figure 4. Phenotypic and molecular characterization of seeds from *lnc1246* and *At3g12940* mutants. A**)** Genomic region showing *AT3G12940* transcribed from the sense strand and *LNCRNA_1246* transcribed from the antisense strand. Mutant allele symbols are as follows: ⊤ is a strand-specific artificial microRNA and ▽ is a T-DNA insertion. B) Strand-specific reverse transcription PCR (RT-PCR) quantifying *LNCRNA_1246* RNA abundance. C) Strand-specific RT-PCR quantifying *AT3G12940* mRNA abundance. *ACTIN1* is an internal control. D) Seed size of *lnc1246* mutants. E**)** Left panel, seed phenotype of *At3g12940* mutant; right panel, mRNA abundance of *AT3G12940* in the mutant. F**)** Weight of wild type and *lnc1246* mutant seeds. Error bars indicate ±SE of the mean (**p* < 0.05, Student's *t*-test; *n* = 100 seeds). Scale bar in panels D and E is 2.5 mm. All RT-qPCRs were performed on two biological and three technical replicates and a representative PCR is shown.

In addition to our observations of smaller seeds in *lnc1246* mutant plants, we found that all examined sporophytic tissue organs (roots, leaves, petals, carpels, anther filaments and siliques) were smaller than in the wild type (Figure 5A–F). We observed no difference in leaf number at flowering (Figure 5E).

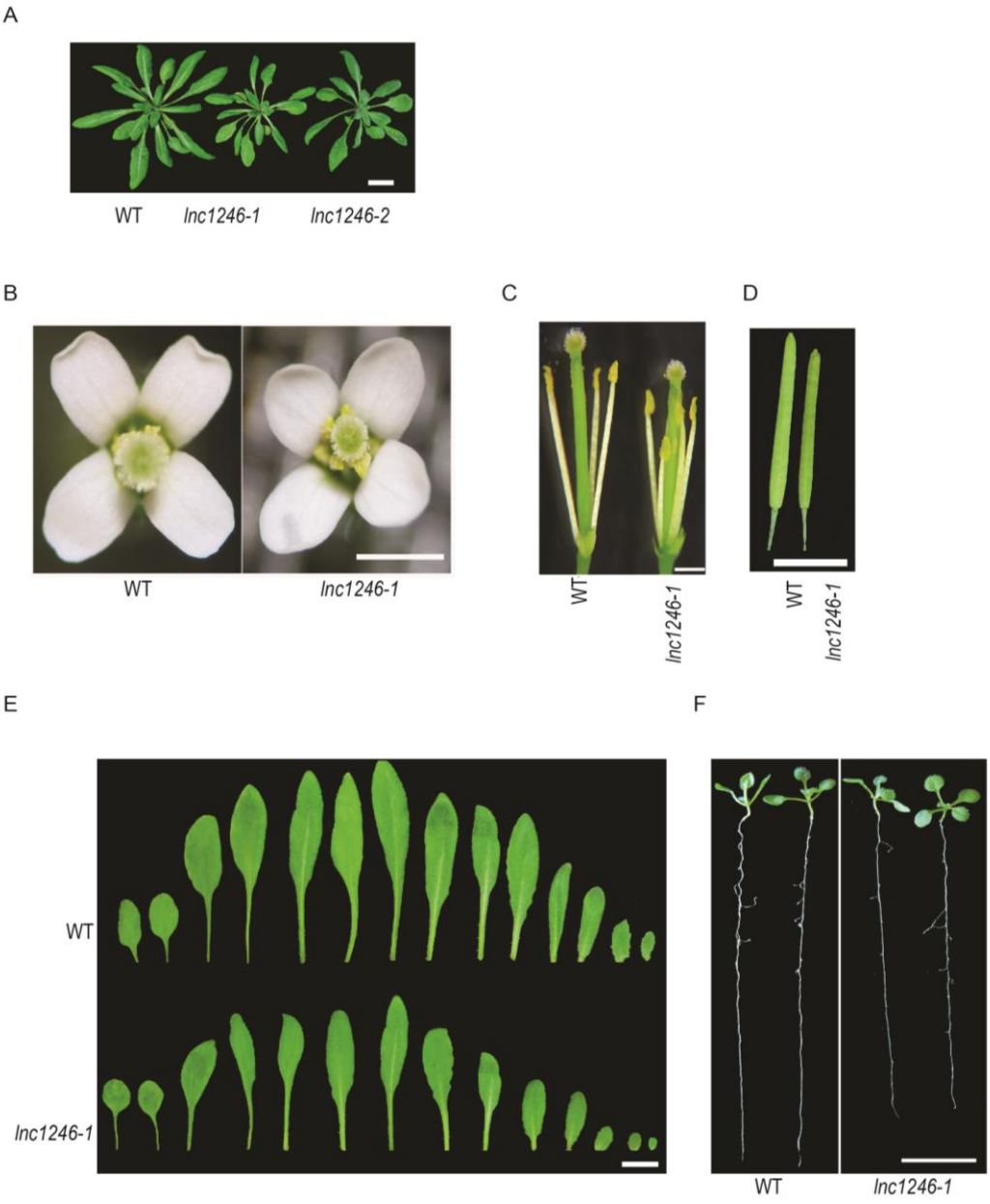


Figure 5. Sporophytic tissues are smaller in long non-coding RNA *lnc1246* mutant plants. A) Vegetative phase; B) flowers; C), stamens and capels; D) siliques 10 days after pollination; E) rosette leaves; and F) roots of seedlings of wild type and *lnc1246* mutant plants. Scale bars panels are as follows: A, 2 cm; B and C, 1 mm; D, E and F, 1 cm.

***LNCRNA_1246* is important for cell expansion in the integuments**

Having observed that mature seeds of *lnc1246* mutants are smaller than wild type, we asked whether the mature cotyledons were also smaller in the mutant. We germinated seeds of wild type and the mutant on plates and observed cotyledons (Figure 6). Consistently, the cotyledons of *lnc1246* mutants were smaller than those of wild type (Figure 6A & 6B). Next, we observed the epidermal cell size in cotyledons using light microscopy (Figure 6C). Epidermal cotyledon cell size was smaller in *lnc1246* mutant seedlings than in wild type. In addition to the epidermal cell layer, we also observed cell size in the sub-epidermal cell layers; they were also smaller than in wild type (see Figure 6S in Appendices).

During *Arabidopsis* seed development, the ovule is surrounded by the integument that develops into the seed coat after fertilization. The effect of smaller integuments leading to smaller seed size has been previously reported [5, 22-23]. Hence, we tested the effect of *lnc1246* mutant integuments on seed size. First, we characterized the mature ovules from wild type and *lnc1246* mutants at 2 days after emasculation. The ovules of *lnc1246* mutant plants were significantly smaller than those of wild type (Figure 6D upper panel). We also observed the outer integument cell size and number, 1 day after fertilization (Figure 6D lower panel, E). The outer integument cell size was smaller in the mutant than in wild type (Figure 6E). Interestingly, outer integument cell number was greater in the mutant compared with wild type, suggesting the existence of a compensation mechanism (Figure 6E). These results show that *LNCRNA_1246* has a role in cell expansion in the integuments of developing seeds.

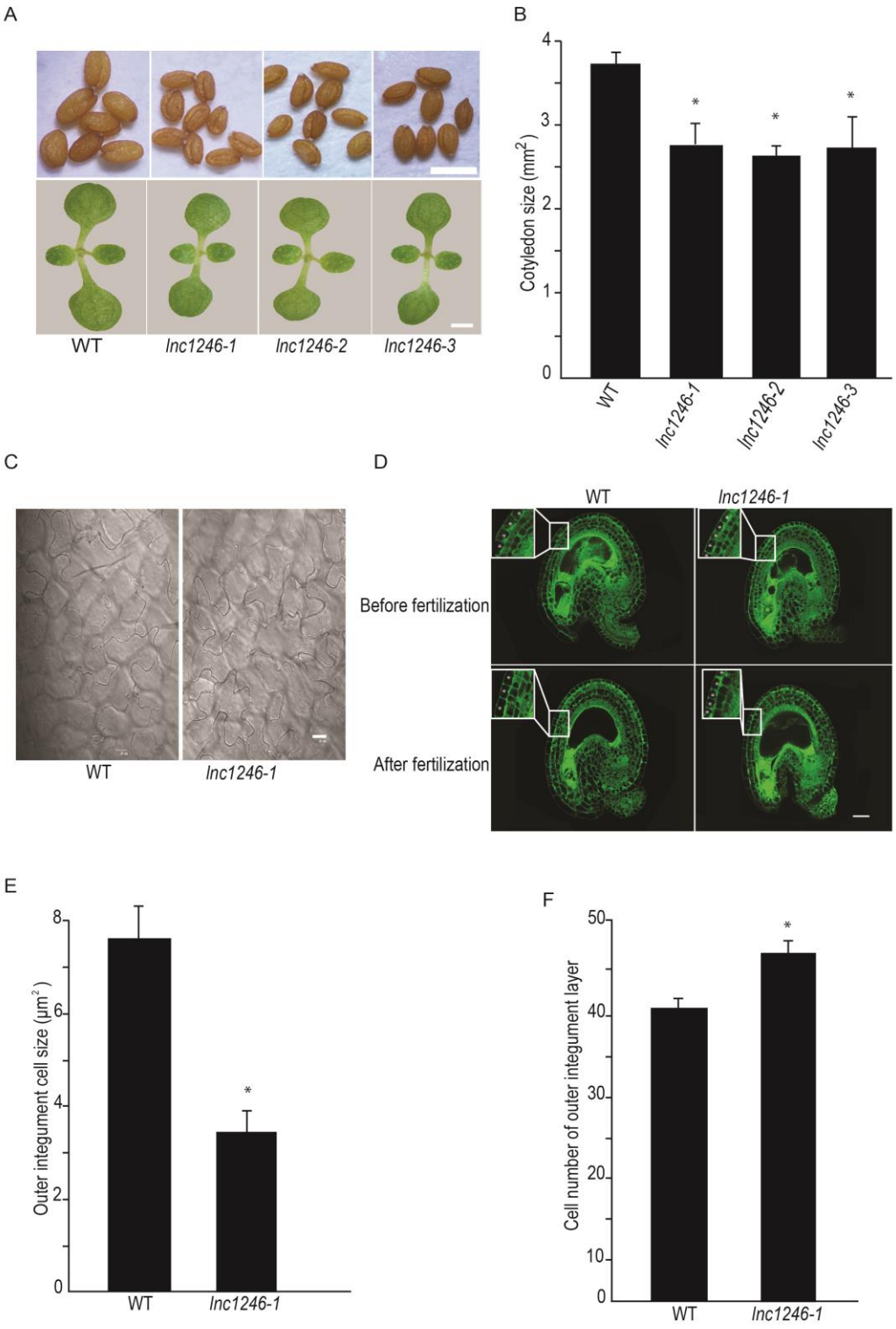


Figure 6. *LNCRNA_1246* is an important regulator of cell size. A) Mature seeds and 10-day-old seedlings from wild type and *lnc1246* mutants (*lnc1246-1, lnc1246-2, lnc1246-3*). B) Quantitfication of cotyledon size of wild type and *lnc1246* mutant seedlings. C) Epidermal cotyledon cells of wild type and *lnc1246-1* seedlings. D) Confocal images of ovules of wild type and *lnc1246* mutants before (upper panel) and after pollination (lower panel). The upper-left insert

in each panel shows an enlarged image of the white box. The cell number is highlighted as white stars. E) The area of cells in the outer integument at 1 day after pollination (DAP). F) The number of cells in the outer integument at 1 DAP. Values in B, D and E are given as means ± SEs (*p < 0.05, Student's t-test; n ≥ 8 for seedlings). Scale bars: A, 0.25 mm for seeds and 1 mm for cotyledons; C, 20 µm; D, 5 µm.

LNCRNA_1246 has a maternal effect on seed size

Maternal and paternal factors have an effect on seed size [6, 22]. Hence, we asked whether *LNCRNA_1246* has a maternal or paternal effect on seed size. To address this, reciprocal crosses between wild type and the dominant *lnc1246-1* mutant were performed. Interestingly, only when *lnc1246-1* was used as a maternal parent were seeds smaller (Figure 7A). Seeds from wild type crossed with wild type and wild type crossed with *lnc1246-1* were the same size (Figure 7A). Together these data suggest that the maternal integuments have a large effect on seed size.

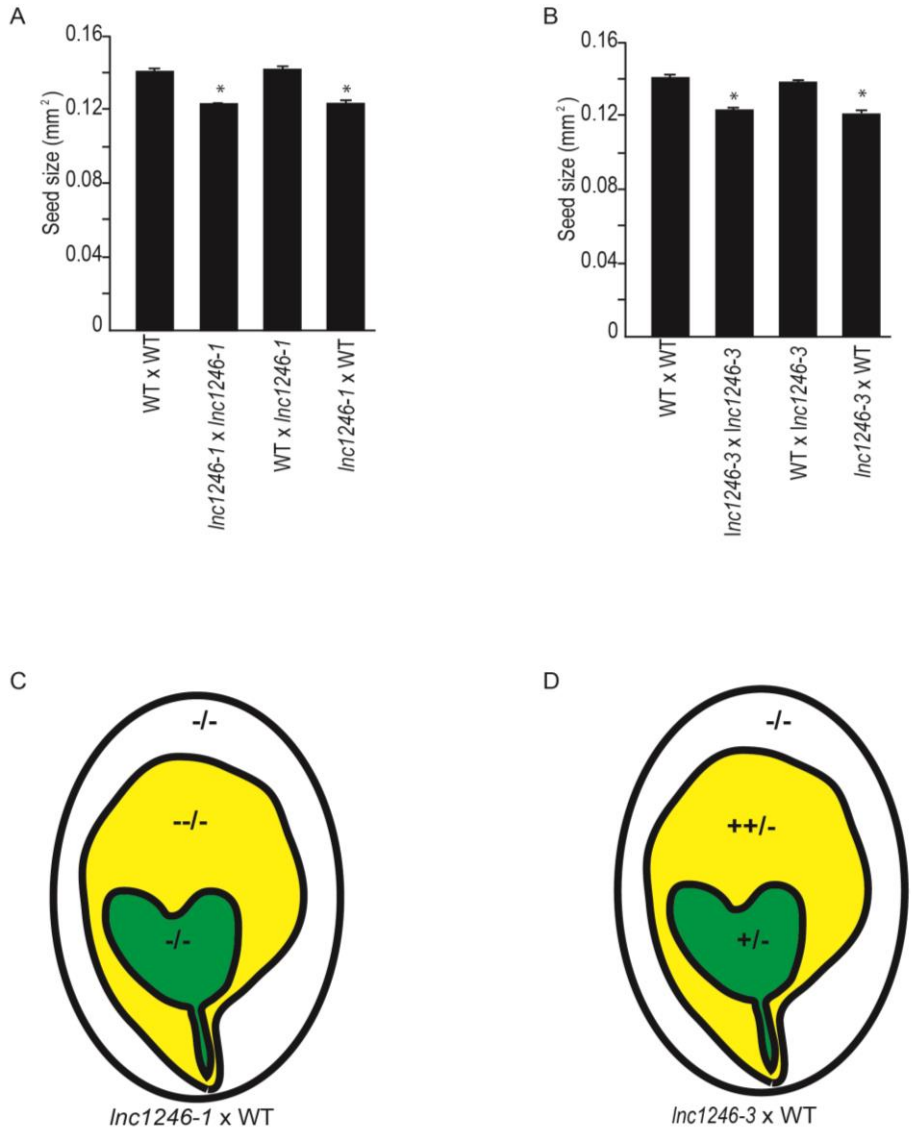


Figure 7. *LNCRNA_1246* acts maternally to regulate seed size. A) Seed size of F_1 seeds from reciprocal crosses between *lnc1246* mutants and wild type. The *lnc1246* mutant allele is a dominantly acting artificial microRNA. B) Seed size of F_1 seeds from reciprocal crosses between *lnc1246* T-DNA mutant and wild type. The *lnc1246* T-DNA mutant allele is a recessively inherited allele. C) and D) Cartoons indicating the genotypes present in different compartments of the F_1 seed from crosses in A and B. Error bars in A and B indicate the SE of the mean (*$p < 0.05$, Student's *t*-test; $n \geq 50$ seeds).

As the *lnc1246* amiRNA parent is dominant, we could not discriminate maternal or paternal effects acting in the endosperm or embryo on seed size (Figure 7C & 7D). Therefore, we performed reciprocal crosses between wild type and the *lnc1246-3* recessive T-DNA insertion allele. Pollinating the *lnc1246-3* maternal parent with wild type pollen produced seeds that were heterozygous for *lnc1246*

in the embryo and endosperm within a mutant seed coat (Figure 7D). We observed that F_1 seeds from these crosses were smaller than self-pollinated wild type seeds. In addition, the seeds from wild type pollinated with *lnc1246-3* mutant pollen were of similar size to the seeds from self-pollinated wild type seeds (Figure 7B). Together, these results suggest that the genotype of *LNCRNA_1246* in the embryo and endosperm does not affect seed size, and that *LNCRNA_1246* acts maternally to regulate seed size.

LNCRNA_1246 may act independently of known genes affecting seed size

Protein-coding genes that act in the maternal integument to promote cell proliferation or cell expansion have been previously described [4–8, 23]. Our results suggest that the smaller seeds of *lnc1246* mutants are the result of reduced cell expansion in the integument before and after fertilization, although we cannot rule out paternal imprinting of *LNCRNA_1246* in the embryo and/or endosperm. We asked whether the mRNA abundance of genes, when mutated to reduce seed size, was reduced in the *lnc1246-1* mutant seed, using RT-qPCR. Our results showed that the mRNA abundance of all the tested genes (*MINI3*, *HAIKU1*, *KLUH, HAIKU2* and *TTG2*) was the same in the mutant and wild type. These results suggest that *LNCRNA_1246* functions separately to these genes to control seed size.

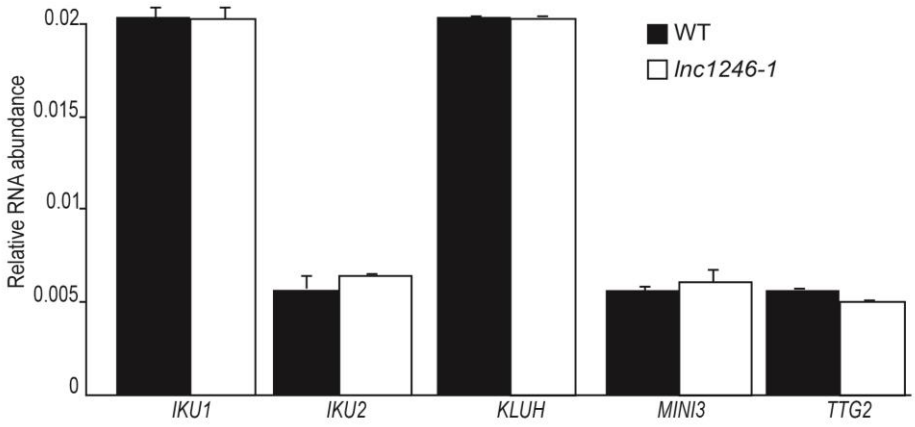


Figure 8. mRNA abundance of important genes regulating seed size in *lnc1246-1* mutant and wild type seeds. RT-qPCR results for *IKU1*, *IKU2*, *KLUH*, *MINI3* and *TTG2* in wild type and *lnc1246-*

1 mutant seeds. RT-qPCRs were performed on two biological and three technical replicates. Error bars indicate ±SE of the mean.

LNCRNA_1246 is not associated with the FIS–PRC2 complex

Next, we tested whether *LNCRNA_1246* was associated with the FIS–PRC2 complex in the endosperm of *A. thaliana* seeds. To address this, we used RNA from the immunoprecipitation experiment described in Trung Do et al. (in prep) and performed RT-PCR. We did not expect *LNCRNA_1246* to be associated with the complex as the sequence was absent from our Illumina sequencing data (Figure 9). We performed RT-qPCR and as expected did not detect *LNCRNA_1246* associated with the FIS–PRC2 complex (Figure 9). However, we did detect by RT-qPCR the positive control lncRNA_29066 that was present in our sequencing data (Figure 9).

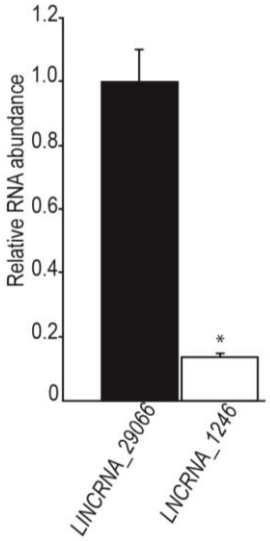


Figure 9. *LNCRNA_1246* is not associated with the FIS–PRC2 complex. RT-PCR results quantifyng assocaition of LINCRNA-29066 [Trung Do et al., in prep] or *LNCRNA_1246* with FIS2–PRC2 after RNA immunoprecipitation. RT-qPCR was performed on two biological and three technical replicates. Error bars indicate ±SE of the mean. *p*-values were calculated with a Student's *t*-test. The asterisk denotes $p < 0.05$.

Transcript characterization of *LNCRNA_1246*

LncRNAs can be broadly classified as two types: those with a polyA-3' end and those without. Hence, we asked whether *LNCRNA_1246* has a 3'-polyA tail. To

address, we carried out 3' RACE (Rapid Amplification of cDNA 3’ Ends) assays for *LNCRNA_1246*; however we could not PCR amplify *LNCRNA_1246* (data not shown here). As a control, we detected *LNCRNA_1246* after strand-specifically priming cDNA synthesis (data not shown here).

Next, we predicted the *LNCRNA_1246* RNA secondary structure based on pairing probability of nucleotides in the sequence using RNAfold [24]. The result is shown in Figure 10. The *LNCRNA_1246* had high free energy (dG = –291.20 kcal mol^{-1}) and several stable stem loops, suggesting a strong secondary structure.

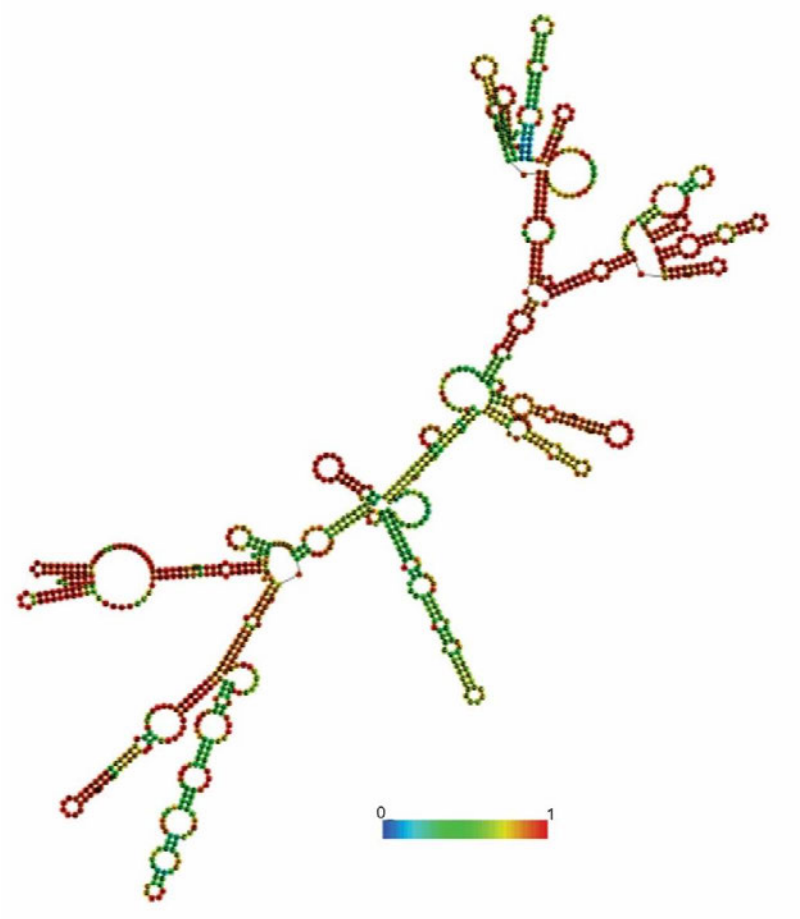


Figure 10. Predicted secondary structure of *LNCRNA_1246* as calculated by RNAfold. The predicted structure is coloured as base-pairing probabilities. Red indicates high and blue indicates low nucleotide pairing probability.

Coding potential of *LNCRNA_1246*

LncRNAs are defined as long RNA molecules with no protein-coding potential and little or no sequence similarity to protein-coding genes. However, several annotated lncRNAs have been associated with ribosomes in both animals [25] and plants [26], suggesting that lncRNAs may also be translated into proteins and hence are bifunctional. Therefore, we predicted potential ORFs in *LNCRNA_1246* using ORF Finder [27]. Twenty-seven potential ORFs were predicted in *LNCRNA_1246* but only one ORF of 33 amino acids started with a methionine (Figure 11A).

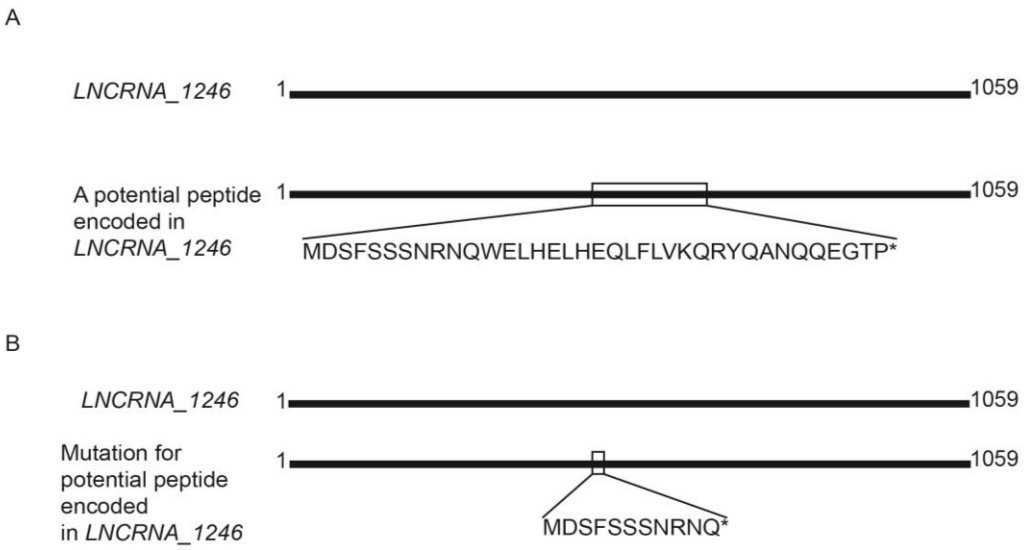


Figure 11. *LNCRNA_1246* putative open reading frame (ORF) analysis. A) ORF1 potentially has 33 amino acids. B) A single nucleotide polymorphism, G to A, was generated to mutate the potential ORF.

To determine if the potential ORF in *LNCRNA_1246* complements the cell size phenotype of *lnc1246* mutants, we mutated the peptide sequence by changing the tryptophan codon (TGG, W) into a stop codon (TGA, *) (Figure 11B). This mutant construction driven by the endogenous promoter was transformed into the recessive *lnc1246-1* mutant background but the phenoptyes of transgenic plants are still to be characterized.

Discussion

With the advantages of biotechnology combined with advances in next generation techniques for genome-wide mapping, genome-wide identification of

lncRNAs has been reported for some plants, including strawberry [28], tomato [29] and *Arabidopsis* [17]. To identify functional roles of lncRNAs in plant endosperm development, we applied a purification protocol for nuclei from early stages of endosperm development of *Arabidopsis thaliana* and sequenced the RNAs. We identified the transcriptome of *A. thaliana* endosperm and experimentally identified lncRNAs associated with seed development. These data will be extremely useful for other researchers and for functional genomics studies and regulatory expression research in the future.

Plant lncRNAs are not well conserved during evolution

The discovered *A. thaliana* endosperm-associated lncRNAs have some distinct characteristics compared with protein-coding genes; for example, low conservation in comparison with currently known lncRNAs from different plant species (data not shown), lower levels of expression, fewer exons and shorter transcripts lengths. These features are shared with plant lncRNAs identified in other studies [28-32]. Most (85%) of the identified *A. thaliana* endosperm-associated lncRNAs had only one exon; this might be due to the choice of parameters during the filtration of novel transcripts, which does not include the number of exons. However, our RT-PCR results showed that all of the single-exon lncRNA candidates gave products; hence, we may lose some real lncRNAs if we remove single-exon transcripts. In addition, the RT-PCR analysis demonstrated the tissue-specific expression of many lncRNAs (Figure 3B). This result is consistent with previous studies that reported that lncRNA spatiotemporal expression profiles are highly tissue-specific [30-31]. The low conservation of the identified lncRNAs compared with currently known lncRNAs from different plant species (data not shown) means that most of the identified lncRNAs were not well conserved and may undergo rapid evolution. Similar results have been reported for lncRNAs from other plants such as tomato [29], maize [32] and *Populus* spp. [30]. This low conservation has several potential explanations: (1) plant lncRNA databases are still in progress; (2) plant species may have their own specific mechanisms to adapt to the environment during evolution and lncRNAs might contain short conserved motifs that are not easily identified in BLAST searches [33] or small interfering RNA encoded by lncRNAs

may be less constrained in other parts of transcripts [34-35]; (3) changes to important factors may affect the formation of a large family of gene homologues, for example, lncRNAs may interact directly with FIS2–PRC2 through its conserved secondary structure [31, 34]; and (4) TEs might play a major role in the generation of alternative promoters, and hence of novel lncRNAs [17, 36-39]. Moreover, the consistency of the RNA-seq and RT-qPCR results provides further evidence that our prediction accuracy was sufficient.

Endosperm lncRNA might controll *Arabidopsis* seed development

The major current challenges in exploring lncRNA function include that (1) they do not encode proteins, hence we cannot apply methods used to analyze protein-coding genes; and (2) they are expressed at only low levels. Recent studies of the tissue-specific expression patterns of lncRNAs have shown that these patterns might help reveal the potential functions of lncRNAs [28]. In this study, we applied a bioinformatics pipeline to identify 615 endosperm-associated lncRNAs that were confirmed by RT-PCR to be endosperm-specifically expressed (Figure 3A). In addition, our data show that the presence of endosperm-associated *LNCRNA_1246* influences reproductive development in *A. thaliana*. This result is consistent with reports in which the tissue-specific regulation of lncRNAs has revealed critical functions during reproductive development in plants and animals [33]. Therefore, our finding provides more evidence for the specific expression of lncRNAs and also suggests that the tissue specificity of lncRNAs is correlated with organ development.

***LNCRNA_1246* controls seed development by regulating maternal integument cell size**

Plant growth and development processes depend on both environmental and endogenous signals that play important roles in determining the anatomy, physiology and molecular features of the plant. Among endogenous signals, lncRNAs have been reported in some plant species as regulatory factors in biological processes such as root developmental plasticity, regulation of phosphate homeostasis, flowering and response to stress [17, 31, 40]. In this study, we found that knockdown of *LNCRNA_1246* produced a smaller seed

phenotype, which might be the result of smaller cell size in the outer integument and not primarily caused by its effect on fertility. This result is consistent with previous reports of the effect of vegetative stage on silique development in *Arabidopsis* [30, 41].

The results from our reciprocal crosses indicate that *LNCRNA_1246* maternally affects seed growth by regulating ovule size via a reduction in integument cell size. The integument is one part of the developing ovule, which is maternal tissue and will form the seed coat after fertilization [9]. Our results show that the size of the ovule in a *lnc1246* mutant is smaller than that of wild type both before and after pollination, suggesting that the size difference arises via alteration of the maternal integument size. This result is consistent with results from other studies in which the integument has been reported to play a role in changing seed size by alternation of its size [5–6]. Hence, our results support the general theme that the maternal integument plays a critical role in determining final seed size.

The molecular roles of *LNCRNA_1246* during plant organogenesis

The molecular explanation for regulating seed size through action *LNCRNA_1246* is still not clear. Our results indicate similar abundances of genes affecting seed size in wild type and mutant plants, suggesting that there might be no interaction between *LNCRNA_1246* and those genes to control seed size (Figure 8). In addition, our results revealed a nuclear localization for *LNCRNA_1246* (Figure 3B) and that it does not bind to the FIS2–PRC2 complex (Figure 9). In animals, lncRNAs located in the nucleus have been shown to play important roles in regulating gene expression at the transcriptional level via histone or DNA modification [42]. Hence, one possible molecular function of *LNCRNA_1246* might be an epigenetic regulatory function during dosage compensation, imprinting or developmental gene expression [43]. The molecular mechanisms by which lncRNAs carry out their functions in this biological process require further study.

In addition to organ development, cell proliferation and cell expansion are two cellular processes that have been shown to have important roles in determining the overall organ size [44-46]. Recently, some protein-coding genes have been

reported to play a role in regulating cell expansion and hence affecting final organ size in *Arabidopsis*, including *EXPANSIN10 (EXP10)*, *REGULATORY PARTICLE AAA-ATPASE 2a (RPT2a)*, *ARGOS-LIKE (ARL)*, *TARGET OF RAPAMYCIN (TOR)*, *ErbB-3 EPIDERMAL GROWTH FACTOR RECEPTOR BINDING PROTEIN 1 (EBP1)* and *ORGAN SIZE RELATED 2 (OSR2)* [44]. Here, we showed that the *lnc1246* mutant regulates cell expansion during organ development (Figure 6C & 6D). *LNCRNA_1246* is expressed in the nuclei of cells from different organs undergoing cell expansion and its knockdown leads to reduced overall organ size by reducing the cell expansion rate (Figure 3B, 6B & 6E). Therefore, our results suggest that *LNCRNA_1246* might have a role as a regulatory factor in plant organ growth and final cell size.

Although the function of most lncRNAs remains unknown, the discovery of lncRNAs from *Arabidopsis* early development siliques provides additional material for future functional studies to understand the biological roles and regulatory mechanisms of lncRNA function in plants.

Perspectives

Understanding seed development and (epi)genetic controls is becoming increasingly important because of the significant role that seeds play as a food source for humans and livestock, as well as the growing demand for biofuel. With cutting-edge genomics-based research, we have identified novel genes with potential roles in seed development. However, there are many knowledge gaps in the field, such as (1) our limited understanding of the mechanisms and networks that act together; (2) the fact that application of the genetic information from model plants to crop plants remains a major challenge [47-49]; and (3) the fact that most of the genome encodes non-coding RNAs and these have no clear function [34]. Hence, more research is required in model and crop plants to understand and improve seed yield.

References

1. Mizukami Y, Fischer RL. Plant organ size control: *AINTEGUMENTA* regulates growth and cell numbers during organogenesis. Proc Nat Acad Sci USA. 2000; 97:942-7.

2. Garcia D, Saingery V, Chambrier P, Mayer U, Jurgens G, Berger F. Arabidopsis *haiku* mutants reveal new controls of seed size by endosperm. Plant Physiol. 2003;131:1661-70.

3. Jofuku KD, Omidyar PK, Gee Z, Okamuro JK. Control of seed mass and seed yield by the floral homeotic gene *APETALA2*. Proc Natl Acad Sci USA. 2005;102:3117-22.

4. Luo M, Dennis ES, Berger F, Peacock WJ, Chaudhury A. *MINISEED3 (MINI3)*, a WRKY family gene, and *HAIKU2 (IKU2)*, a leucine-rich repeat (LRR) *KINASE* gene are regulators of seed size in Arabidopsis. Proc Nat Acad Sci USA. 2005; 102:17531-6.

5. Schruff MC, Spielman M, Tiwari S, Adams S, Fenby N, Scott RJ. The *AUXIN RESPONSE FACTOR 2* gene of *Arabidopsis* links auxin signaling, cell division, and the size of seeds and other organs. Dev. 2006;133:251-61.

6. Adamski NM, Anastaslou E, Eriksson S, O'Neill CM, Lenhard M. Local maternal control of seed size by *KLUH/CYP78A5*-dependent growth signalling. Proc Nat Acad Sci USA. 2009;106:20115-20.

7. Zhou Y, Zhang X, Kang X, Zhao X, Zhang X, Ni M. SHORT HYPOCOTYL UNDER BLUE1 associates with *MINISEED3* and *HAIKU2* promoters in vivo to regulate *Arabidopsis* seed development. Plant Cell. 2009;21:106-17.

8. Wang A, Garcia D, Zhang H, Feng K, Chaudhury A, Berger F, Peacock WJ, Dennis ES, Luo M. The VQ motif protein IKU1 regulates endosperm growth and seed size in *Arabidopsis*. Plant J. 2010;63:670-9.

9. Berger F, Grini PE, Schnittger A. Endosperm: an integrator of seed growth and development. Curr Opin Plant Biol. 2006;9:664–70.

10. Baroux C, Pien S, Grossniklaus U. Chromatin modification and remodelling during early seed development. Curr Opin Genet Dev. 2007;17:473–9.

11. Kiyosue T, Ohad N, Yadegari R, Hannon M, Dinneny J, Wells D, Katz A, Margossian L, Harada JJ, Goldberg RB, Fischer RL. Control of fertilization-independent endosperm development by the *MEDEA* polycomb gene in *Arabidopsis*. Proc Natl Acad Sci USA. 1999;96:4186-91.

12. Chaudhury AM, Ming L, Miller C, Craig S, Dennis ES, Peacock WJ. Fertilization-independent seed development in *Arabidopsis thaliana*. Proc Natl Acad Sci USA. 1997;94:4223-8.

13. David R, Burgess A, Parker B, Li J, Pulsford K, Sibbritt T, Preiss T, Searle I. Transcriptome-wide mapping of RNA 5-methylcytosine in Arabidopsis mRNAs and ncRNAs. The Plant Cell. 2017:29(3) 445-60.

14. Burgess AL, David R, Searle IR. Conservation of tRNA and rRNA 5-methylcytosine in the kingdom Plantae. BMC Plant Biol. 2015:15:199.

15. Deal RB and Henikoff S. The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana. Nat Protoc. 2011:6(1):56-68.

16. Moreno-Romero J, Santos-González J, Hennig L, Köhler C. Applying the INTACT method to purify endosperm nuclei and to generate parental-specific epigenome profiles. Nat Protoc. 2017:12(2):238-54.

17. Wang D, Qu Z, Yang L, Zhang Q, Liu ZH, Do T, Adelson DL, Wang ZY, Searle I, Zhu JK. Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in Plants. Plant J. 2017:90:133-46.

18. Jakoby M, Wang HY, Reidt W, Weisshaar B, Bauer P FRU (BHLH029) is required for induction of iron mobilization genes in *Arabidopsis thaliana*. FEBS Lett. 2004:19;577(3):528-34.

19. Davis AM, Hall A, Millar AJ, Darrah C, and Davis SJ. Protocol: Streamlined sub-protocols for floral-dip transformation and selection of transformants in *Arabidopsis thaliana*. Plant Methods. 2009:5:3.

20. Bai Y, Dai X, Harrison AP, Chen M. RNA regulatory networks in animals and plants: a long noncoding RNA perspective. Brief Funct Genomics. 2015:14(2):91-101.

21. Tiwari S, Schulz R, Ikeda Y, Dytham L, Bravo J, Mathers L, Spielman M, Guzmán P, Oakey RJ, Kinoshita T, Scott RJ. MATERNALLY EXPRESSED PAB C-TERMINAL, a novel imprinted gene in Arabidopsis, encodes the conserved C-terminal domain of polyadenylate binding proteins. Plant Cell. 2008 Sep;20(9):2387-98.

22. Garcia D, Fitz Gerald JN, Berger F. Maternal control of integument cell elongation and zygotic control of endosperm growth are coordinated to determine seed size in *Arabidopsis*. Plant Cell. 2005;17:52–60.

23. Dilkes BP, Spielman M, Weizbauer R, Watson B, Burkart-Waco D, Scott RJ, Comai L. The maternally expressed WRKY transcription factor *TTG2* controls lethality in interploidy crosses of *Arabidopsis*. PLoS Biol. 2008;6:2707–20.

24. Andreas R. Gruber, Ronny Lorenz, Stephan H. Bernhart, Richard Neuböck and Ivo L. Hofacker. The Vienna RNA Websuite. Nucleic Acids Res. 2008; 36(Web Server issue): W70–W74.

25. Nam JW, Choi SW, You BH. Incredible RNA: Dual Functions of Coding and Noncoding. Mol Cells. 2016:39(5):367–74.

26. Lauressergues D, Couzigou JM, Clemente HS, Martinez Y, Dunand C, Bécard G, Combier JP. Primary transcripts of microRNAs encode regulatory peptides. Nature. 2015: 520(7545):90-3.

27. Stothard P. The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques. 2000:28:1102-1104.

28. Kang C, Liu Z. Global identification and analysis of long noncoding RNAs in diploid strawberry *Fragaria vesca* during flower and fruit development. BMC Genomics. 2015;16:815.

29. Zhu B, Yang Y, Li R, Fu D, Wen L, Luo Y, Zhu H. RNA sequencing and functional analysis implicate the regulatory role of long non-coding RNAs in tomato fruit ripening. Journal of Experimental Botany. 2015;66:4483–95.

30. Wang CY, Liu SR, Zhang XY, Ma YJ, Hu CG, Zhang JZ. Genome-wide screening and characterization of long noncoding RNAs involved in flowering development of trifoliate orange (*Poncirus trifoliata L. Raf*.). Sci Rep. 2017;7:43226.

31. Di C, Yuan J, Wu Y, Li J, Lin H, Hu L, Zhang T, Qi Y, Gerstein MB, Guo Y, Lu ZJ. Characterization of stress-responsive lncRNAs in *Arabidopsis thaliana* by integrating expression, epigenetic and structural features. Plant Journal. 2014;80:848–61.

32. Li L, Eichten SR, Shimizu R, Petsch K, Yeh C-T, Wu W, Chettoor AM, Givan SA, Cole RA, Fowler JE. Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol. 2014;15(2):R40.

33. Golicz AA, Bhalla PL, Singh MB. lncRNAs in Plant and Animal Sexual Reproduction. Trends Plant Sci. 2018::23(3):195-205.

34. Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, Leyva A, Weigel D, García JA, Paz-Ares J. Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet. 2007;39(8):1033–1037.

35. Zhang X, Xia J, Lii YE, Barrera-Figueroa BE, Zhou X, Gao S, Lu L, Niu D, Chen Z, Leung C, Wong T, Zhang H, Guo J, Li Y, Liu R, Liang W, Zhu JK, Zhang W, Jin H. Genomewide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function. Genome Biol. 2012;13(3):R20.

36. Oosumi T, Gruszewski HA, Blischak LA, Baxter AJ, Wadl PA, Shuman JL, Veilleux RE, Shulaev V. High-efficiency transformation of the diploid strawberry (*Fragaria vesca*) for functional genomics. Planta. 2006;223(6):1219–30.

37. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012;13(11):R107.

38. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest AR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009;41(5):563–71.

39. Johnson R, Guigó R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA. 2014;20(7):959–76.

40. Bazin J, Bailey-Serres J. Emerging roles of long non-coding RNA in root developmental plasticity and regulation of phosphate homeostasis. Plant Sci. 2015; 6:400.

41. Tamura K, Kawabayashi T, Shikanai T, Hara-Nishimura I. Decreased Expression of a Gene Caused by a T-DNA Insertion in an Adjacent Gene in *Arabidopsis*. PLoS ONE. 2016;11(2): e0147911.

42. Nakagawa S, Kageyama Y. Nuclear lncRNAs as epigenetic regulators—Beyond scepticism. Biochimica et Biophysica Acta. 2014: 1839: 215–222.

43. Kim D-H, Xi Y, Sung S. Modular function of long noncoding RNA, COLDAIR, in the vernalization response. PloSGenet 2017:13(7): e1006939.

44. Gonzalez N, Vanhaeren H, Inzé D. Leaf size control: complex coordination of cell division and expansion. Trends Plant Sci. 2012; 17(6):332–340.

45. Kalve S, De Vos D, Beemster GT. Leaf development: a cellular perspective. Front Plant Sci. 2014; 5:362.

46. Qin Z, Zhang X, Zhang X, Feng G, Hu Y. The *Arabidopsis* ORGAN SIZE RELATED 2 is involved in regulation of cell expansion during organ growth. BMC Plant Biology. 2014;14:349.

47. Wu HJ, Wang ZM, Wang M, Wang XJ. Widespread long noncoding RNAs as endogenous target mimics for microRNAs in plants. Plant physiology. 2013;161:1875–84.

48. Gheysen G, Herman L, Breyne P, Gielen J, Van Montagu M, Depicker A. Cloning and sequence analysis of truncated T-DNA inserts from *Nicotiana tabacum*. Gene. 1990;94(2):155–63.

49. Nacry P, Camilleri C, Courtial B, Caboche M, Bouchez D. Major chromosomal rearrangements induced by T-DNA transformation in *Arabidopsis*. Genetics. 1998;149(2):641–50.

Chapter 7: General Discussion

7.1 Context of This Study

Genome-wide analysis of the human genome has shown that a substantial proportion of the genome is transcribed into a wide range of RNAs differing in size, abundance and protein-coding capability (Cheng et al., 2005; Kapranov et al., 2005; Carninci et al., 2005; Katayama et al., 2005; Birney et al., 2005; Djebali et al., 2012). Similar observations have recently been made in plants (Liu et al., 2012; Li et al., 2014; Kang and Liu, 2015). However, only a very small proportion of these transcripts are translated into proteins; the majority is untranslated and these are broadly termed ncRNAs. These ncRNAs are crudely divided into small ncRNAs (fewer than 200 nts) and lncRNAs (more than 200 nts) (Ponting et al., 2009). Unlike small ncRNAs, which have been well studied, lncRNAs remain largely poorly characterized, especially in plants (Chitwood and Timmermans, 2010; Liu et al., 2012; Chen et al., 2015). To date, no lncRNAs involved in endosperm or embryo development have been described.

In this thesis, we describe in detail a methodology for purification of specific cell types, bioinformatics annotation of lncRNAs and investigation of biological function using the reference plant *A. thaliana* (Chapter 2). We also detail methods for highly reproducible bisulfite treatment of RNA, efficient locus-specific PCR amplification, detection of candidate sites by sequencing on the Illumina MiSeq platform and bioinformatic calling of converted and non-converted sites (Chapter 3).

To investigate the role of TE-derived lncRNAs in plant development, we identified and characterized TE-associated lincRNAs (TE-lincRNAs) from *Arabidopsis*, rice and maize (Chapter 4). TEs have been reported as major contributors to the origin, diversification and regulation of lncRNAs from human, mouse, zebrafish and recently tomato (Kapusta et al., 2013; Wang et al., 2016). Here we showed that TEs make a contribution to the origin of stress-related lincRNAs from *Arabidopsis*, rice and maize. Using loss-of-function mutants, we demonstrated a role for some TE-lincRNAs under stress, but not control conditions. This suggests that TE-lincRNAs may act as an adaptive reservoir in eukaryotes.

To identify lncRNAs involved in epigenetic regulation of seed development, we sequenced whole seeds and enriched for lncRNAs bound to the important FIS2–PRC2 complex (Chapter 5). The FIS2–PRC2 complex is important in regulating seed development and acts to restrain endosperm proliferation before and after fertilization by depositing repressive H3K27me3 histone marks on target genes (Weinhofer et al., 2010; Butenko and Ohad, 2011). How the complex is recruited to target genes is largely unknown and two somewhat opposing models have been proposed: the DNA transcription factor and ncRNA guide models. Here it was demonstrated that thousands of lncRNAs are bound to the FIS2–PRC2 complex and may function in regulating target genes. Interestingly, the data also showed that G-tract motifs (G2L1-4) are significantly enriched among PRC2-binding transcripts. While incomplete, these data support a *cis*-acting model whereby lncRNAS regulate PRC2 complex activity.

To investigate the function of FIS2–PRC2-bound lncRNAs, we identified and characterized T-DNA insertion mutants for 9 PRC2-associated lncRNAs and characterized the expression of neighbouring genes, which was previously reported to be upregulated in *fis2* mutants (Chapter 5). Using qRT-PCR, we found that mutation of FIS2–PRC2-associated lncRNAs show a strong association with nascent H3K27me3 target genes. This suggests a possible mechanism for predicted lncRNAs in regulating their neighbour genes by binding and guiding the FIS2–PRC2 to the target sites.

To explore the function of lncRNAs during endosperm development, we used a methodology for lncRNA purification of specific cell types (Chapter 1). We identified 615 lncRNAs in *A. thaliana* endosperm nuclei (Chapter 6). We showed that these lncRNAs have tissue-specific expression as many exhibit a relative abundance difference among tissue types or are unique to one tissue type, suggesting tissue-specific functions and regulation. Of those novel lncRNAs that are common to multiple tissue types, some are differentially expressed among tissue types, while others have the same level of expression across all tissue types. Further, we experimentally demonstrated that knockdown of *LNCRNA_1246* results in a decrease in seed size by reducing the cell expansion of outer integument cells. Through reciprocal crosses I demonstrated that

LNCRNA_1246 acts maternally to regulate seed size. In addition to seed development, we also showed that *LNCRNA_1246* is a general regulator of cell and organ size. In summary, the results support the function of lncRNA in seed development.

Together these data represent a transcriptome-wide, high-resolution view of lncRNAs in *A.* thaliana, rice and maize and in association with the FIS2–PRC2 complex, and provide links to biological function. In addition, the identification and characterization of *Arabidopsis* loss-of-function genetic mutants provides a valuable resource for the research community to further build upon in detail to establish the functions of these lncRNAs in the future. Substantial discussion of these findings has already been presented in Chapters 4, 5 and 6. In the following sections, I discuss the broader implications of RNA regulatory networks in animals and plants, with a focus on lncRNAs.

7.2 RNA Regulatory Networks in the Evolution of Animals and Plants

Before the current genomics era (ca. 2000 onwards), it was suggested that the number of protein-coding genes that an organism made use of was a valid measure of its complexity. However, it is now clear that there is only a weak relationship between biological complexity and the number of protein-coding genes. Further, the proteomes of higher organisms are relatively stable. For example, mice and humans have 99% of their protein-coding genes in common. It is now clear that very few nucleotide polymorphisms between phenotypically different individuals within higher organisms reside in protein-coding regions and similar observations have been made between related species. Thus, phenotypic variation among individuals and among related species may be based largely on differences in non-protein-coding nucleotide sequences. Very recently, it has become clear that most (>95%) transcription from higher organism genomes is non-protein-coding RNA. At the beginning of this study, there were only limited reports of the extent of lncRNAs in plants and fewer functional characterization reports.

Recent genomics research has discovered many families of transcripts that have function but do not code for proteins; termed ncRNAs. An important group of ncRNAs is lncRNAs, whose members originate from thousands of loci across the genome. There is growing evidence that lncRNAs can regulate gene expression at the transcriptional and post-transcriptional levels and take part in various physiological and pathological processes, such as cell development, immunity, oncogenesis and clinical disease processes, among others. All of this evidence suggests a central role of lncRNAs as master regulators of gene expression and chromatin organization that might make them particularly suited for coordination and control of molecular processes involved in animal and plant evolution. Here, I discuss why lncRNAs could be a central player in the evolution of animals and plants, in three sections: 1) Diversity of lncRNAs—Substrates for plant and animal evolution; 2) Regulatory function—Emerging role of lncRNAs; and 3) Evolution of lncRNAs.

7.2.1 Diversity of long noncoding RNAs—Substrates for plant and animal evolution

Transcriptome studies have shown that more than 75% of the human genome is actively transcribed into protein-coding transcripts (mRNAs) and ncRNAs (Cheng et al., 2005; Kapranov et al., 2005; Carninci et al., 2005; Katayama et al., 2005; Birney et al., 2005). Interestingly, the proportion of mRNAs is very small and widespread occurrence of ncRNAs has been demonstrated (Wu et al., 2017).

LncRNAs can be subdivided into several classes based on their relationship to protein-coding genes and different mechanisms of processing. In relation to protein-coding genes, different classes of lncRNA transcripts—such as promoter upstream transcripts (PROMPTs), enhancer RNAs (eRNAs), lincRNAs and natural antisense transcripts (NATs) have been transcribed from promoter upstream regions, enhancers, intergenic regions and the opposite strand of protein-coding genes in eukaryotic genomes (Wu et al., 2017). In addition, many new lncRNA species with unexpected structures are generated from long primary transcripts with unusual RNA-processing pathways (Wu et al., 2017). For example, instead of using canonical 5'-end m^7G capping or 3'-end poly (A) tailing for maturation, lncRNAs can be stabilized by several non-canonical mechanisms,

including RNase P cleavage to generate a mature 3' end (Wilusz et al., 2008; Sunwoo et al., 2009); capping by small nucleolar ribonucleoproteins (snoRNPs) at both ends (Yin et al., 2012; Zhang et al., 2014b; Xing et al., 2017) or the 5' end (Wu et al., 2016); or forming circular structures to protect them from degradation (Salzman et al., 2012; Jeck et al., 2013; Memczak et al., 2013; Salzman et al., 2013; Zhang et al., 2013b; Zhang et al., 2014c). Notably, many lncRNAs are alternatively spliced to generate multiple isoforms leading to higher diversity of lncRNAs (Johnsson et al., 2013). The data presented here (Chapter 4, 5 and 6) show that thousands of lncRNAs are transcribed from different loci in the genome of *A. thaliana*. In addition, the data demonstrate some distinct features of lncRNAs compared with protein-coding genes: for example, they are on average shorter, show less sequence and positional conservation and are less abundant, but have more tissue-specific expression patterns. These results are consistent with previous reports involving lncRNAs (Derrien et al., 2012; Pauli et al., 2012; Liu et al., 2012; Li et al., 2014; Hezroni et al., 2015; Wang et al., 2015; Khemka et al., 2016). Interestingly, it was also found here that by altering the nuclear chromatin state, new lincRNAs can be generated (Chapter 4). Overall, these discoveries indicate further layers of complexity to gene expression and regulation.

It is believed that the diversity of lncRNA families offers functional diversity in regulatory networks (Lee, 2012; Kung et al., 2013). They can act in *cis*-acting mode to influence neighbouring loci and *trans*-acting mode to perform distal regulatory functions. These modes suggest greater diversity in lncRNA function, in which the *trans*-acting molecules possibly act within larger co-expression networks and *cis*-acting counterparts have more localized roles (Herriges et al., 2014; Alam et al., 2014; Melé et al., 2016; Luo et al., 2016). Although the dominant role of lncRNAs in *cis* or *trans* regulation is still debated, lncRNAs have emerged as new functional molecules found in many eukaryotic forms of life (Amaral and Mattick, 2008; Morris and Mattick, 2014; Golicz et al., 2018). In the next section, I provide more detail about the potential role of lncRNAs in adaptation to changing environments.

7.2.2 Regulatory function—Emerging roles of long noncoding RNAs

Recent discoveries have led to an emerging picture of an extremely rich landscape of diverse RNAs that are transcribed from many loci of the genome in a spatiotemporally dependent manner (Amaral and Mattick, 2008; Morris and Mattick, 2014; Golicz et al., 2018). An important group of regulatory ncRNAs are lncRNAs, which may play an important role in the adaptation of plants and animals to a changing environment (Amaral and Mattick, 2008; Golicz et al., 2018). I now discuss in more detail the regulatory function of lncRNAs as an important part of the epigenetic landscape that controls differentiation and development in plants and animals.

Most functionally analyzed lncRNAs seem to play a role in regulating differentiation and development in plants and animals (Amaral and Mattick, 2008; Golicz et al., 2018). It has been reported that many lncRNAs have functions as master regulators of gene expression and chromatin organization involved in sexual reproduction of plants and animals (Golicz et al., 2018). However, this may be an oversimplification of their function in adaptive processes. Interestingly, the data in this thesis shows that knockdown of *LNCRNA_1246* led to smaller cell and organ phenotypes in all tested tissues including roots, cotyledons and seeds (Chapter 6). In addition to the developmental role of lncRNAs, I also presented results that knockdown of TE-associated lincRNA_11195 caused reduced sensitivity to ABA by producing longer and thicker roots compared with wild-type plants after ABA treatment (Chapter 4). These data may suggest a broader role for lncRNAs under stress or adaptation to stress.

Consistent with their roles in differentiation and development, the huge amount of evidence from genetic and biochemical studies demonstrates important functions of lncRNAs in epigenetic regulation by guiding chromatin-modifying enzymes to their target sites and/or acting as scaffolds for chromosomal organization (Mattick et al., 2009; Khalil et al., 2009; Mercer and Mattick, 2013). For example, some naturally occurring lncRNAs have been shown to control epigenetic processes such as X chromosome dosage compensation and parental imprinting in mammals (Sado et al., 2001; Thakur et al., 2004), and vernalization in plants (Swiezewski et al., 2009). Subsequent studies have shown that

antisense and intergenic lncRNAs bind to PRCs to alter chromatin modification and/or DNA methylation, leading to allele-specific silencing (see Table 2 in Chapter 1). The results from the RIP experiment in this study identified thousands of lncRNAs bound to FIS2_PRC2 in *A. thaliana* (Chapter 5). While the functions of most of these PRC2-associated transcripts is unknown, I showed that knockdown of some PRC2-associated lncRNAs by T-DNA insertion resulted in gene expression changes and the upregulation of nearby genes that were normally silenced by PRC2 (Chapter 5). In addition to epigenetic regulation, I found that lincRNAs could play a role in the alteration of chromatin state by regulating DNA methylation on chromosomes (Chapter 4). These data suggest lncRNAs play an important role in the epigenetic processes that control the differentiation and development of plants and animals.

Regulatory lncRNA expression may be influenced by environmental signals and transmitted between cells and even generations, which could be important in the evolution of plants and animals. For example, flowering time regulation in *A. thaliana* depends on cold conditions that trigger expression of COLDAIR (lncRNA) leading to vernalization (Swiezewski et al., 2009). The results here showed that the expression of TE-lincRNAs in *A. thaliana* was affected by different stress conditions (Chapter 4). Further, the results indicated that *ddm1* produced many unique lincRNAs that may also play a role in responses to stress. These may contribute to the biotic stress resistance found in *ddm1* (Dowen et al., 2012) and were interestingly inherited in a wild-type background in subsequent generations (Chapter 4). These data demonstrate that regulatory lncRNAs play important roles in plant stress responses.

Overall, it is becoming clear that lncRNAs are an important part of the regulatory networks in plants and animals. Importantly, the evolution of lncRNAs in response to environmental signals over generations puts them in an important position in plant and animal evolution.

7.2.3 Evolution of long noncoding RNAs

RNA is thought to have played a variety of important roles in the evolution of life on the earth. Many important discoveries have revealed that regulatory RNAs

play important roles in the diversification of life, which has resulted in a string of innovations by RNA (Amaral and Mattick, 2008; Bai et al., 2014; Golicz et al., 2018). An important group is the lncRNAs that have been co-opted into the regulatory systems of plants and animals (Bai et al., 2014; Golicz et al., 2018). Here, I discuss in more detail the evolution of lncRNAs in terms of low sequence conservation, and TEs as contributors to lncRNAs.

Nucleotide sequence conservation is often useful for predicting function of coding and sometimes noncoding genes (Cooper and Brown, 2008; Kellis et al., 2014). Many studies have attempted to measure functional constraints on lncRNA exon sequences within and across species for animals (Guttman et al., 2009; Marques and Ponting, 2009; Young et al., 2012) and recently plants (Derrien et al., 2012; Liao et al., 2014; Zhang et al., 2014a; Song et al., 2016). Not surprisingly, conservation analysis shows low sequence conservation of lncRNAs in both plants and animals, suggesting rapid turnover of lncRNAs, in contrast to the evolutionary nucleotide stability of protein-coding genes (Kapusta and Feschotte, 2014). The conservation analysis performed here also indicated low conservation of lncRNAs among different species (Chapters 4, 5 and 6). Although the correlation between sequence conservation and expression is positive for both lncRNAs and protein-coding genes, lncRNAs seem to be more sensitive to changes in expression levels than are protein-coding genes (Managadze et al., 2011; Popadin et al., 2013; Nielsen et al., 2014; Wang et al., 2017). This is consistent with the current results, in which expression of selected TE-associated lncRNAs changed more rapidly than that of neighbouring genes under different stress conditions (Chapter 4). Further, in animals, lncRNAs show a rapid decrease of orthologous expression conservation during evolution relative to their sequence conservation, while the orthologous mRNA expression level is much more consistent across mammals (Washietl et al., 2014). Together, these data suggest that lncRNA expression level is more prone to change during evolution than is that of mRNAs (Necsulea et al., 2014).

Interestingly, the data revealed that some TE-associated lncRNAs have a role in abiotic responses (Chapter 4). This result suggests that insertion of TEs may contribute to evolution and function of ncRNAs in two ways. First, two-thirds of

mammalian genomes and a significant proportion of plant genomes are TEs, which can be grouped into various classes—retroelements, endogenous retroviruses, DNA transposons and so on—and hundreds of families (Gregory, 2005; de Koning et al., 2011). TEs could provide the raw material to assemble or modify genetic function. For example, TEs have been reported to play important roles in genome structure and to provide the material for evolution (such as new protein-coding genes, transcription factor binding sites and connecting gene regulatory networks) because of their capability to move and amplify and the ability to introduce new regulatory sequences after insertion (Feschotte, 2008; Villar et al., 2014). Thus, it can be proposed that the significant activity of TEs during evolution serves as a source of hypermutagenicity that could create useful diversity among individuals in a population (Feschotte, 2008; Villar et al., 2014).

Second, recent discoveries have shown that TEs can also contribute to the origin and diversity of lncRNAs (Kapusta et al., 2013; Wang et al., 2016; Chishima et al., 2018). TEs are commonly observed within mature lncRNAs in both vertebrates and plants. It has been estimated that two-thirds of vertebrate lncRNAs (Kapusta et al., 2013) and more than 20% of *A. thaliana* lincRNAs (Chapter 4) contain at least one TE-derived sequence, whereas TEs are rarely present in protein-coding genes. It was shown here that TEs contribute to new lincRNAs under stress conditions. Interestingly, by altering the chromatin state in the mutant *ddm1*, we showed that unique lincRNAs were inherited in a subsequent wild-type background, suggesting that the *de novo* evolution from sequences derived from TEs might account for the birth of new lincRNAs (Chapter 4). These results also provide evidence for the rapid emergence of lncRNAs from TEs and their interaction in regulatory networks controlling development. Collectively the data demonstrate the profusion and diversity of TEs embedded into lncRNAs and shows that their interactions with the cell machinery are promiscuous, complex and modular (Goodier and Kazazian, 2008; Levin and Moran, 2011; Wang et al., 2017; Chishima et al., 2018). These data suggest an important contribution of TEs to the evolution and diversity of lncRNAs.

In summary, lncRNA genes would evolve very differently from protein-coding genes in response to environmental signals, and are transmitted between cells and generations. Interestingly, during evolution, a vast number of lncRNAs have been generated but they still retain the unique features of lncRNAs that distinguish them from protein-coding genes. All of these factors guarantee them a role in the evolution of plants and animals.

7.3 Conclusions and Future Directions

In summary, this thesis presents the transcriptome-wide identification of lncRNAs from different species (maize, rice and *A. thaliana* [seedling and endosperm]) and demonstrates the first transcriptome of lncRNAs that is associated with the FIS2–PRC2 complex or derived from TEs in *A. thaliana*. These lncRNAs are tissue-specifically regulated. Moreover, I report the identification and characterization of lncRNA mutants affecting stress response and cell size. This provides a means to further investigate the functions of these lncRNAs in plant development.

Future experiments examining the expression pattern of *LNCRNA_1246*, *in vivo* protein–lncRNA interactions and the effect of TE-associated lncRNAs on seed development will establish a more detailed view of how *LNCRNA_1246* may act.

In the future, these data, combined with other recent findings uncovering the *Arabidopsis* epitranscriptome will be important for understanding complex biological phenomena such as hybrid vigour, stress responses and hybridization barriers.

References Cited

Adamski, NM, Anastasiou, E, Eriksson S, O'Neill, CM, and Lenhard, M 2009, "Local maternal control of seed size by KLUH/CYP78A5-dependent

growth signalling", *Proceedings of the National Academy of Sciences of the United States of America*, vol.106(47), pp.20115-20120.

Allan, RL and Abed, C 2002, "Genetic and epigenetic processes in seed development", *Current Opinion in Plant Biology*, vol. 5, pp.19–25.

Alam, T, Medvedeva, YA, Jia, H, Brown, JB, Lipovich, L, Bajic, VB 2014, "Promoter analysis reveals globally differential regulation of human long noncoding RNA and proteincoding genes", *PLoS One*, vol.9, e109443.

Amaral, P P & Mattick, J S, 2008, "Noncoding RNA in development", *Mamm. Genome*, vol.19, pp.454–492.

Ariel, F, Jegu, T, Latrasse, D, Romero-Barrios, N, Christ, A, Benhamed, M, Crespi, M, 2014, "Noncoding transcription by alternative RNA polymerases dynamically regulates an auxin-driven chromatin loop", *Mol Cell*, *vol*. 55(3), pp.383–396.

Bai, Y, Dai, X, Harrison, PA and Chen, M 2014, "RNA regulatory networks in animals and plants: a long noncoding RNA perspective", *Briefings in Functional Genomics,* elu017v1-elu017.

Bailey-Serres, J, Sorenson, R and Juntawong, P 2009, "Getting the message across: cytoplasmic ribonucleoprotein complexes", *Trends Plant Sci*, vol.14, pp.443–453.

Bardou, F, Ariel, F, Simpson, CG, Romero-Barrios, N, Laporte, P, Balzergue, S, Brown, JWS, Crespi, M, 2014, "Long noncoding RNA modulates alternative splicing regulators in *Arabidopsis*", *Dev Cell*, *vol.* 30(2), pp.166–176.

Baud, S, Wuillème, S, Lemoine, R, Kronenberger, J, Caboche, M, Lepiniec, L and Rochat, C 2005, "The AtSUC5 sucrose transporter specifically expressed in the endosperm is involved in early seed development in *Arabidopsis*", *The Plant Journal : for Cell and Molecular Biology*, vol.43(6), pp.824-836.

Baud, S, Dubreucq, B, Miquel, M, Rochat, C, Lepiniec, L 2008, "Storage reserve accumulation in *Arabidopsis*: metabolic and developmental control of seed filling", *Arabidopsis* Book. Vol.6:e0113.

Beddington, J 2010, "Food security: contributions from science to a new and greener revolution", *Philos.Trans.R.Soc.London*, Ser.B365, pp.61–71.

Berger, F, 1999, "Endosperm development", *Curr. Opin. Plant Biol*., vol.2, pp.28–32.

Berger, F 2003, “Endosperm: the crossroad of seed development”, *Curr. Opin. Plant Biol*. Vol. 6(1), pp. 42-50.

Beisel, C and Paro, R 2011, “Silencing chromatin: comparing modes and mechanisms”, *Nature Reviews Genetics*, vol. 12, pp.123–135.

Birney, E, Stamatoyannopoulos, JA, Dutta, A, Guigó, R, Gingeras, TR, Margulies, EH, Weng, Z et al 2007, “Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project”, *Nature*, vol.447, pp.799–816.

Borges, F, Parent, JS, van, Ex F, Wolff, P, Martínez, G, Köhler, C, Martienssen, RA 2018, “Transposon-derived small RNAs triggered by miR845 mediate genome dosage response in *Arabidopsis*”, *Nat Genet.*, vol.50(2), pp.186-192.

Böhmdorfer, G, Sethuraman, S, Rowley, MJ, Krzyszton, M, Rothi, MH, Bouzit, L, Wierzbicki, AT, 2016, “Long noncoding RNA produced by RNA polymerase V determines boundaries of heterochromatin”, *eLife*, *vol.* 5, pp.1325.

Boisnard-Lorig, C, Colon-Carmona, A, Bauch, M, Hodge, S, Doerner, P, Bancharel, E, Dumas, C, Haseloff, J and Berger, F 2001, “Dynamic analyzes of the expression of the HISTONE::YFP fusion protein in *Arabidopsis* show that syncytial endosperm is divided in mitotic domains”, *Plant Cell*, vol.13, pp.495-509.

Butenko, Y and Ohad, N 2011, “Polycomb-group mediated epigenetic mechanism through plant evolution”, *Biochimica et Biophysica Acta,* vol. 1809, pp. 395-406.

Carninci, P, Kasukawa, T, Katayama, S, Gough, J, Frith, MC, Maeda, N, Oyama, R, Ravasi, T, Lenhard, B, Wells, C, et al 2005, “The transcriptional landscape of the mammalian genome”, *Science*, vol.309, pp.1559–1563.

Castello, A, Fischer, B, Eichelbaum, K, Horos, R, Beckmann, BM, Strein, C, Davey, NE, Humphreys, DT, Preiss, T, Steinmetz, LM et al. 2012, “Insights into RNA biology from an atlas of mammalian mRNA-binding proteins”, *Cell*, vol.149, pp.1393–1406.

Canales, C, Bhatt, AM, Scott, R and Dickinson, H 2002, “EXS, a putative LRR receptor kinase, regulates male germline cell number and tapetal identity and promotes seed development in *Arabidopsis*”, *Current biology : CB*, vol.12(20), pp.1718-1727.

Chaudhury, AM, Ming, L, Miller, C, Craig, S, Dennis, ES, Peacock, WJ 1997, “Fertilization-independent seed development in *Arabidopsis thaliana*”, *Proc Natl Acad Sci USA*, vol.94(8), pp.4223-4228.

Chanvivattana, Y, Bishopp, A, Schubert, D, Stock, C, Moon, YH, Sung, ZR, Goodrich, J 2004, “Interaction of polycomb-group proteins controlling flowering in *Arabidopsis*”, *Development*, vol.131, pp.5263-5276.

Cheng, L, Shafiq, S, Xu, W, and Sun, Q 2018, “EARLY FLOWERING IN SHORT DAYS (EFS) regulates the seed size in *Arabidopsis*”, *Sci China Life Sci*, vol.61, pp.214–224.

Cheng, J, Kapranov, P, Drenkow, J, Dike, S, Brubaker, S, Patel, S, Long, J, Stern, D, Tammana, H, Helt, G, Sementchenko, V, Piccolboni, A, Bekiranov, S, Bailey, DK 2005, “Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution”, *Science*, vol.308, pp.1149–1154.

Chen, J, Quan, M and Zhang, D 2015, “Genome-wide identification of novel long noncoding RNAs in *Populus tomentosa* tension wood, opposite wood and normal wood xylem by RNA-sEq”, *Planta*, vol. 241, pp. 125–143.

Chishima, T, Iwakiri, J, Hamada, M 2018, “Identification of Transposable Elements Contributing to Tissue-Specific Expression of Long Noncoding RNAs”, *Genes* (Basel), Vol. 9(1), pp.23.

Chitwood, DH and Timmermans, MC 2010, “Small RNAs are on the move”, *Nature,* vol. 467, pp. 415–419.

Choi, Y, Gehring, M, Johnson, L, Hannon, M, Harada, JJ, Goldberg, RB, Jacobsen, SE, and Fischer, RL 2002, “DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in *Arabidopsis*”, *Cell*, vol.110(1), pp.33-42.

Chu, C, Qu, K, Zhong, FL, Artandi, SE, Chang, HY, 2011, “Genomic maps of long noncoding RNA occupancy reveal principles of RNA–chromatin interactions”, *Mol Cell*, vol. 44, pp.667–678.

Chu, C, Zhang, QC, Da Rocha, ST, Flynn, RA, Bharadwaj, M, Calabrese, JM, Magnuson, T, Heard, E, Chang, HY 2015, “Systematic discovery of Xist RNA binding proteins”, *Cell*, vol. 161, pp.404–416.

Costa, LM, Gutierrez-Marcos, JF and Dickinson, HG 2004, “More than a yolk: the short life and complex times of the plant endosperm”, *Trends Plant Sci*., vol.9, pp.507-514.

Costa, LM, Marshall, E, Tesfaye, M, Silverstein, KA, Mori, M, Umetsu, Y, Otterbach, SL, Papareddy, R, Dickinson, HG, Boutiller, K, VandenBosch, KA, Ohki, S, Gutierrez-Marcos, JF 2014, “Central cell-derived peptides regulate early embryo patterning in flowering plants”, *Science*, vol.344, pp.168–172.

Cooper, GM and Brown, CD 2008, “Qualifying the relationship between sequence conservation and molecular function”, *Genome Res.*, vol. 18, pp. 201–205.

Czech, B and Hannon, GJ 2011, “Small RNA sorting: matchmaking for argonautes”, *Nat. Rev. Genet.*, vol.12, pp. 19–31.

Davidovich, C and Cech, TR 2015, “The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2”, *RNA*, vol.21(12), pp. 2007-2022.

De Lucia, F, Crevillen, P, Jones, AM, Greb, T, Dean, C 2008, “A PHD-polycomb repressive complex 2 triggers the epigenetic silencing of *FLC* during vernalization”, *Proc Natl Acad Sci USA*, vol.105, pp.16831-16836.

de Koning, APJ, Gu, W, Castoe, TA, Batzer, MA, Pollock, DD 2011, “Repetitive Elements May Comprise Over Two-Thirds of the Human Genome”, *PLoS Genet*, vol. 7(12): e1002384.

Derrien, T, Johnson, R, Bussotti, G, Tanzer, A, Djebali, S, Tilgner, H, Guernec, G, Martin, D, Merkel, A and Knowles, DG 2012, “The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression”, *Genome Res.*, vol. 22, pp. 1775–1789.

Djebali, S, Davis, CA, Merkel, A, Dobin, A, Lassmann, T, Mortazavi, A, Tanzer, A, Lagarde, J, Lin, W, Schlesinger, F et al. 2012, “Landscape of transcription in human cells”, *Nature*, vol. 489, pp. 101–108

Dowen, RH, Pelizzola, M, Schmitz, RJ, Lister, R, Dowen, JM, Nery, JR, Dixon, JE and Ecker, JR 2012, “Widespread dynamic DNA methylation in response to biotic stress”, *Proc. Natl Acad. Sci. USA*, vol.109, E2183–E2191.

Fabian, MR, Sonenberg, N and Filipowicz, W 2010, “Regulation of mRNA translation and stability by microRNAs”, *Annu. Rev. Biochem.*, vol.79, pp.351–379.

Farrona, S, Coupland, G and Turck, F 2008, “The impact of chromatin regulation on the floral transition”, *Semin Cell Dev Biol*, vol.19(6), pp.560-573.

Feschotte, C 2008, “Transposable elements and the evolution of regulatory networks”, *Nat. Rev. Genet.*, vol. 9, pp. 397–405.

Feil, R and Berger, F 2007, “Convergent evolution of genomic imprinting in plants and mammals”, *Trends Genet*., vol.23, pp.192–199.

Garcia-Hernandez, M and Reiser, L 2002, “Using information from public *Arabidopsis* databases to aid research”, *Methods in Molecular Biology*, vol. 323, pp.187-211.

Garcia, D, Fitz-Gerald, JN and Berger, F 2005, “Maternal control of integument cell elongation and zygotic control of endosperm growth are coordinated to determine seed size in *Arabidopsis*”, *The Plant Cell,* vol.17(1), pp.52-60.

Gehring, M, Bubb, KL, Henikoff, S 2009, “Extensive demethylation of repetitive elements during seed development underlies gene imprinting”, *Science*, vol.324, pp.1447–1451.

Grini, PE, Jurgens, G and Hulskamp, M 2002, “Embryo and endosperm development is disrupted in the female gametophytic capulet mutants of *Arabidopsis*”, *Genetics*, vol.162, pp.1911–1925.

Gregory, TR 2005, “Synergy between sequence and size in large-scale genomics”, *Nat. Rev. Genet.,* vol. 6, pp. 699–708.

Goodier, JL and Kazazian, HH 2008, “Retrotransposons revisited: the restraint and rehabilitation of parasites”, *Cell*, vol. 135, pp. 23–35.

Golicz, AA, Bhalla, PL, Singh, MB 2018, “lncRNAs in Plant and Animal Sexual Reproduction”, *Trends Plant Sci.*, vol.23(3), pp.195-205.

Grote, P, Wittler, L, Hendrix, D, Koch, F, Währisch, S, Beisaw, A, Macura, K, Bläss, G, Kellis, M, Werber, M 2013, “The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse”, *Dev Cell*, vol. 24, pp.206–214.

Guttman, M, Amit, I, Garber, M, French, C, Lin, MF, Feldser, D, Huarte, M, Zuk, O et al. 2009, “Chromatin signature reveals over a thousand highly conserved large noncoding RNAs in mammals”, *Nature*, vol. 458, pp. 223–227.

Hamamura, Y, Nagahara, S, Higashiyama, T 2012, “Double fertilization on the move”, *Curr Opin Plant Biol.*, vol. 15(1), pp.70-77.

Hacisuleyman, E, Goff, LA, Trapnell, C, Williams, A, Henao-Mejia, J, Sun, L, McClanahan, P, Hendrickson, DG, Sauvageau, M, Kelley, DR 2014,

"Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre", *Nat Struct Mol Biol,* vol. 21, pp.198–206.

Harada, JJ, Belmonte, MF, and Kwong, RW 2010, "Plant Embryogenesis (Zygotic and Somatic)". In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester. DOI: 10.1002/9780470015902.a0002042.pub2.

Haig, D and Westoby, M 1989, "Parent specific gene expression and the triploid endosperm", *Am Nature*, vol.134, pp.147–155.

Hehenberger, E, Kradolfer, D, Köhler, C 2012, "Endosperm cellularization defines an important developmental transition for embryo development", *Development*, vol.139, pp.2031–2039.

Hezroni, H, Koppstein, D, Schwartz, MG, Avrutin, A, Bartel, DP, Ulitsky, I 2015, "Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species", *Cell Rep.*, vol.11, pp.1110–1122.

He, C, Huang, H, Xu, L 2013, "Mechanisms guiding Polycomb activities during gene silencing in *Arabidopsis* thaliana", *Front Plant Sci.* vol.13(4):454.

Heo, JB and Sung, S 2011, "Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA", *Science*, vol.331, pp.76–79.

Heo, JB, Lee, YS, Sung, S 2013, "Epigenetic regulation by long noncoding RNAs in plants", *Chromosome Res*, vol. 21(6-7), pp. 685-693.

Herriges, MJ, Swarr, DT, Morley, MP, Rathi, KS, Peng, T, Stewart, KM, Morrisey, EE 2014, "Long noncoding RNAs are spatially correlated with transcription factors and regulate lung development", *Genes Dev.*, vol.28, pp.1363–1379.

Hsieh, TF, Ibarra, CA, Silva, P, Zemach, A, Eshed-Williams, L, Fischer, RL and Zilberman, D 2009, "Genome-wide demethylation of *Arabidopsis* endosperm", *Science*, vol.324, pp.1451–1454.

Hsieh, TF, Shin, J, Uzawa, R, Silva, P, Cohen, S, Bauer, MJ, Hashimoto, M, Kirkbride, RC, Harada, JJ, Zilberman, D, Fischer, RL 2011, "Inaugural article: regulation of imprinted gene expression in *Arabidopsis* endosperm", *Proc. Natl. Acad. Sci. USA*, vol.108, pp.1755–1762.

Huarte, M, Guttman, M, Feldser, D, Garber, M, Koziol, MJ, Kenzelmann-Broz, D, Khalil, AM, Zuk, O, Amit, I, Rabani, M 2010, "A large intergenic noncoding

RNA induced by p53 mediates global gene repression in the p53 response". *Cell*, vol.142, pp.409–419.

Hung, T, Wang, Y, Lin, MF, Koegel, AK, Kotake, Y et al. 2011, "Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters", *Nat. Genet.*, vol.43, pp.621–629.

Jeck, WR, Sorrentino, JA, Wang, K, Slevin, MK, Burd, CE, Liu, J, Marzluff, WF, Sharpless, NE 2013, "Circular RNAs are abundant, conserved, and associated with ALU repeats", *RNA*, vol.19, pp.141–157.

Jenik, PD, Gillmor, S and Lukowitz, W 2007, "Embryonic patterning in *Arabidopsis thaliana*", *Annu. Rev. Cell Dev. Biol*., vol.23, pp. 207-236.

Johnsson, P, Ackley, A, Vidarsdottir, L, Lui, WO, Corcoran, M, Grandér, D, Morris, KV 2013, "A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells", *Nature Struct. Mol. Biol.*, vol.20, pp.440–446.

Jullien, PE, Kinoshita, T, Ohad, N, Berger, F 2006, "Maintenance of DNA methylation during the *Arabidopsis* life cycle is essential for parental imprinting", *Plant Cell*, vol.18, pp.1360–1372.

Kapusta, A, Kronenberg, Z, Lynch, VJ, Zhuo, X, Ramsay, L, Bourque, G, Yandell, M, Feschotte, C 2013, "Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs", *PLoS Genet.*, vol. 9(4): e1003470.

Kapusta, A and Feschotte, C 2014, "Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications", *Trends in Genetics*, Vol. 30(10), pp. 439-452.

Kang, IH, Steffen, JG, Portereiko, MF, Lloyd, A and Drews, GN 2008, "The AGL62 MADS domain protein regulates cellularization during endosperm development in *Arabidopsis*", *The Plant Cell*, vol.20 (3), pp.635-647.

Kang, X, Li, W, Zhou, Y, Ni, M 2013, "A WRKY transcription factor recruits the SYG1-like protein SHB1 to activate gene expression and seed cavity enlargement", *PLoS Genet*, vol.9:e1003347.

Kang, C and Liu, Z 2015, "Global identification and analysis of long noncoding RNAs in diploid strawberry *Fragaria vesca* during flower and fruit development", *BMC Genomics*, vol.16(1):815.

Kapranov, P, Drenkow, J, Cheng, J, Long, J, Helt, G, Dike, S, Gingeras, TR 2005, "Examples of the complex architecture of the human transcriptome

revealed by RACE and high-density tiling arrays", *Genome Res.*, vol.15, pp.987–997.

Katayama, S, Tomaru, Y, Kasukawa, T, Waki, K, Nakanishi, M, Nakamura, M, Nishida, H, Yap, CC, Suzuki, M, Kawai, J, et al. 2005, "Antisense transcription in the mammalian transcriptome", *Science*, vol.309, pp.1564–1566.

Kannan, S, Chernikova, D, Rogozin, IB, Poliakov, E, Managadze, D, Koonin, EV, Milanesi, L 2015, "Transposable Element Insertions in Long Intergenic Noncoding RNA Genes", *Front. Bioeng. Biotechnol.*, Vol.*3*, pp.71.

Kellis, M, Wold, B, Snyder, MP, Bernstein, BE, Kundaje, A, Marinov, GK, Ward, LD, Birney, E, Crawford, GE, et al. 2014, "Defining functional DNA elements in the human genome", *Proc. Natl. Acad. Sci.* U.S.A., vol. 111, pp. 6131–6138.

Kelley, D and Rinn, J 2012, "Transposable elements reveal a stem cell-specific class of long noncoding RNAs", *Genome Biol.*, vol. 13, R107.

Keene, JD 2007, "RNA regulons: coordination of post-transcriptional events", *Nature Review Genetics, vol.* 8(7), pp. 533–543.

Khemka, N, Singh, VK, Garg R, Jain M 2016, "Genome-wide analysis of long intergenic noncoding RNAs in chickpea and their potential role in flower development", *Sci. Rep.*, vol.6, pp.33297.

Khalil, AM, Guttman, M, Huarte, M, Garber, M, Raj, A, Rivea Morales D, Thomas, K, Presser, A, Bernstein, BE, van Oudenaarden, A, Regev, A, Lander, ES, Rinn, JL 2009, "Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression", *Proc. Natl Acad. Sci. USA*, vol.106, pp.11667–11672.

Kinoshita, T, Miura, A, Choi, Y, Kinoshita, Y, Cao, X, Jacobsen, SE, Fischer, RL and Kakutani, T 2004, "One-way control of *FWA* imprinting in *Arabidopsis* endosperm by DNA methylation", *Science*, vol.303, pp. 521–523.

Kino, T, Hurt, DE, Ichijo, T, Nader, N and Chrousos, GP 2010, "Noncoding RNA *Gas5* is a growth arrest and starvation-associated repressor of the glucocorticoid receptor", *Sci. Signal*, *vol.*3:ra8.

Kim, DH and Sung, S 2017, "Vernalization-triggered intragenic chromatin loop formation by long noncoding RNAs", *Dev Cell*, *vol.* 40(3), pp.302–312.e4.

Köhler, C, Hennig, L, Bouveret, R, Gheyselinck, J, Grossniklaus, U, Gruissem, W 2003, "*Arabidopsis* MSI1 is a component of the MEA/FIE polycomb

group complex and required for seed development", *EMBO J.*, vol.22, pp.4804-4814.

Köhler, C and Weinhofer-Molisch, I 2010, "Mechanisms and evolution of genomic imprinting in plants", *Heredity*, vol.105, pp.57-63.

Köhler, C and Kradolfer, D 2011, "Epigenetic mechanisms in the endosperm and their consequences for the evolution of flowering plants", *Biochim Biophys Acta*., vol.1809, pp.438-443.

Kondou, Y, Nakazawa, M, Kawashima, M, Ichikawa, T, Yoshizumi, T, Suzuki, K, Ishikawa, A, Koshi, T, Matsui, R, Muto, S and Matsui, M 2008, "RETARDED GROWTH OF EMBRYO1, a new basic helix-loop-helix protein, expresses in endosperm to control embryo growth", *Plant Physiology*, vol.147(4), pp.1924-1935.

Köster, T and Staiger, D 2014, "RNA-Binding Protein Immunoprecipitation from Whole-Cell Extracts", in Sanchez-Serrano, J. J. and Salinas, J., "*Arabidopsis* Protocols", Totowa, NJ, Humana Press, vol.1062, pp.679-695.

Kruszka, K, Pieczynski, M, Windels, D, Bielewicz, D, Jarmolowski, A, Szweykowska-Kulinska, Z, Vazquez, F, 2012, "Role of miroRNAs and other sRNAs of plants in their changing environments", *J Plant Physiol*, vol.169(16), pp.1664-72.

Kunarso, G, Chia, NY, Jeyakani, J, Hwang, C, Lu, X, Chan, YS, Ng, HH, Bourque, G 2010, "Transposable elements have rewired the core regulatory network of human embryonic stem cells", *Nature Genetics*, Vol.42, pp. 631–634.

Kung, JT, Colognori, D, Lee, JT 2013, "Long noncoding RNAs: past, present, and future", *Genetics*, vol.193, pp.651–669.

Lafon-Placette, C and Kohler, C 2014, "Embryo and endosperm, partners in seed development", *Curr Opin Plant Biol*, vol.17, pp.64-69.

Lee, JT 2012, "Epigenetic regulation by long noncoding RNAs", *Science*, vol. 338, pp. 1435-1439.

Leeuwen, S and Mikkers, H 2010, "Long noncoding RNAs: Guardians of development", *Differentiation*, vol. 80, pp. 175–183.

Levin, HL and Moran, JV 2011, "Dynamic interactions between transposable elements and their hosts", *Nat. Rev. Genet.*, vol. 12, pp. 615–627.

Li, Y, Zheng, L, Corke, F, Smith, C, Bevan, MW 2008, “Control of final seed and organ size by the DA1 gene family in *Arabidopsis thaliana*”, *Genes Dev.*, vol.22, pp.1331–1336.

Li, Y, Li, C, Xia, J, Jin, Y 2011, “Domestication of transposable elements into MicroRNA genes in plants”, *PLoS One* 6: e19212.

Li, J and Berger, F 2012, “Endosperm: food for humankind and fodder for scientific discoveries”, *New Phytol*., vol.195 (2), pp.290-305.

Li, L, Eichten, SR, Shimizu, R, Petsch, K, Yeh, CT, Wu, W, Chettoor, AM, Givan, SA, Cole, RA, Fowler, JE, Evans, MM, Scanlon, MJ, Yu, J, Schnable, PS, Timmermans, MC, Springer, NM, Muehlbauer, GJ 2014, “Genome-wide discovery and characterization of maize long noncoding RNAs”, *Genome Biol.*, vol.15, pp.R40.

Liao, Q, Shen, J, Liu, J, Sun, X, Zhao, G, Chang, Y, Xu, L, Li, X, Zhao, Y, Zheng, H 2014, “Genome-wide identification and functional annotation of Plasmodium falciparum long noncoding RNAs from RNA-seq data”, *Parasitol. Res.*, vol. 113, pp. 1269–1281.

Liu, F, Quesada, V, Crevillen, P, Baurle, I, Swiezewski, S and Dean, C 2007, “The *Arabidopsis* RNA-binding protein FCA requires a lysine-specific demethylase 1 homolog to downregulate *FLC*”, *Mol. Cell*, vol. 28, pp. 398–407.

Liu, F, Marquardt, S, Lister, C, Swiezewski, S and Dean, C 2010, “Targeted 3’ processing of antisense transcripts triggers *Arabidopsis FLC* chromatin silencing”, *Science*, vol. 327, pp. 94–97.

Liu, X, Kim, YJ, Muller, R, Yumul, RE, Liu, C, Pan, Y, Cao, X, Goodrich, J, Chen, X 2011, “AGAMOUS terminates floral stem cell maintenance in *Arabidopsis* by directly repressing *Wuschel* through recruitment of polycomb group proteins”, *Plant Cell*, vol. 23, pp. 3654–3670.

Liu, J, Jung, C, Xu, J, Wang, H, Deng, S, Bernad, L, Arenas-Huertero, C, Chua, NH 2012, “Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*”, *Plant Cell*, vol. 24, pp. 4333–4345.

Lukong, KE, Chang, KW, Khandjian, EW, Richard S, 2008, “RNA-binding proteins in human genetic disease”, *Trends in Genetics, vol.* 24(8), pp. 416–425.

Luo, M, Dennis, ES, Berger, F, Peacock, WJ and Chaudhury, A 2005, “MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a

leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in *Arabidopsis*", *Proceedings of the National Academy of Sciences of the United States of America*, vol.102(48), pp.17531-17536.

Luo, S, Lu, JY, Liu, L, Yin, Y, Chen, C, Han, X, Wu, B, Xu, R, Liu, W, Yan, P, Shao, W, Lu, Z, Li, H, Na, J, Tang, F, Wang, J, Zhang, YE, Shen, X 2016, "Divergent lncRNAs regulate gene expression and lineage differentiation in pluripotent cells", *Cell Stem Cell*, vol.18, pp.637–652.

Lorkovic, ZJ and Barta, A 2002, "Genome analysis: RNA recognition motif (RRM) and K-homology (KH) domain RNA-binding proteins from the flowering plant *Arabidopsis thaliana*", *Nucleic Acids Res*., vol. 30, pp. 623–635.

Marques, AC and Ponting, CP 2009, "Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness", *Genome Biol.*, vol. 10, R124.

Managadze, D, Rogozin, IB, Chernikova, D, Shabalina, SA, Koonin, EV, "Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs", *Genome Biol. Evol.*, vol. 3, pp. 1390–1404.

Martinez, G, Wolff, P, Wang, Z, Moreno-Romero, J, Santos-González, J, Conze, LL, DeFraia, C, Slotkin, RK, Köhler, C 2018, "Paternal easiRNAs regulate parental genome dosage in *Arabidopsis*", *Nat Genet.*, vol.50(2), pp.193-198.

Mattick, JS, Amaral, PP, Dinger, ME, Mercer, TR, Mehler, MF 2009, "RNA regulation of epigenetic processes", *Bioessays*, vol.31, pp.51–59.

Makarevich, G, Villar, CB, Erilova, A, Köhler, C 2008, "Mechanism of PHERES1 imprinting in *Arabidopsis*", *J. Cell Sci.*, vol. 121, pp. 906–912.

McHugh, CA, Chen, C, Chow, A, Surka, CF, Tran, C 2015, "The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3", *Nature*, vol. 521, pp.232–236.

Mercer, TR, Dinger, ME, Mattick, JS 2009, "Long noncoding RNAs: insights into functions", *Nat Rev Genet*, vol. 10, pp. 155–159.

Mercer, TR and Mattick, JS 2013, "Structure and function of long noncoding RNAs in epigenetic regulation", *Nature Struct. Mol. Biol.*, vol.20, pp.300–307.

Memczak, S, Jens, M, Elefsinioti, A, Torti, F, Krueger, J, Rybak, A, Maier, L, Mackowiak, SD, Gregersen, LH, Munschauer, M, Loewer, A, Ziebold, U,

Landthaler, M, Kocks, C, le Noble, F, Rajewsky, N 2013, “Circular RNAs are a large class of animal RNAs with regulatory potency”, *Nature*, vol.495, pp.333–338.

Melé, M, Mattioli, K, Mallard, W, Shechner, DM, Gerhardinger, C, Rinn, JL 2016, “Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs”, *Genome Res.*, vol.27, pp.27–37.

Mili, S and Steitz, JA 2004, “Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyzes”, *RNA*, vol. 10, pp. 1692–1694.

Miller, MT, Higgin, JJ and Tanaka Hall, TM 2008, “Basis of altered RNA-binding specificity by PUF proteins revealed by crystal structures of yeast Puf4p”, *Nat. Struct. Mol. Biol.*, vol. 15, pp. 397–402.

Minic, Z, Do CT, Rihouey, C, Morin, H, Lerouge, P and Jouanin, L 2006, “Purification, functional characterization, cloning, and identification of mutants of a seed-specific arabinan hydrolase in *Arabidopsis*”, *Journal of Experimental Botany*, vol. 57(10), pp. 2339-2351.

Mizukami, Y and Fischer, RL 2000, “Plant organ size control: AINTEGUMENTA regulates growth and cell numbers during organogenesis”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97(2), pp.942-947.

Morris, AR, Mukherjee, N and Keene, JD 2010, “Systematic analysis of posttranscriptional gene expression”, *Wiley Interdiscip. Rev. Syst. Biol. Med.*, vol. 2, pp. 162–180.

Morris, KV and Mattick JS 2014, “The rise of regulatory RNA”, *Nat Rev Genet.*, vol.15(6), pp.423-437.

Moore, MJ 2005, “From birth to death: the complex lives of eukaryotic mRNAs”, *Science*, vol. 309, pp. 1514–1518.

Mosher, RA and Melnyk, CW 2010, “siRNAs and DNA methylation: seedy epigenetics”, *Trends Plant Sci*, vol. 15, pp. 204–210.

Morley-Smith, ER, Pike, MJ, Findlay, K, Köckenberger, W, Hill, LM, Smith, AM, Rawsthorne, S 2008, “The transport of sugars to developing embryos is not via the bulk endosperm in oilseed rape seeds”, *Plant Physiol.*, vol.147(4), pp.2121-2130.

Nagano, T, Mitchell, JA, Sanz, LA, Pauler, FM, Ferguson-Smith, AC, Feil, R, Fraser, P, 2008, “The Air noncoding RNA epigenetically silences

transcription by targeting G9a to chromatin", *Science*, vol. 322, pp.1717–20.

Nesi, N, Debeaujon, I, Jond, C, Stewart, AJ, Jenkins, GI, Caboche, M and Lepiniec L 2002, "The TRANSPARENT TESTA16 locus encodes the *ARABIDOPSIS* BSISTER MADS domain protein and is required for proper development and pigmentation of the seed coat", *The Plant Cell*, vol. 14(10), pp. 2463-2479.

Necsulea, A, Soumillon, M, Warneforts, M, Liechti, A, Daish, T, Zeller, U, Baker JC, Grützner, F, Kaessmann, H, 2014, "The evolution of lncRNA repertoires and expression patterns in tetrapods", *Nature*, vol. 505, pp. 635–640.

Nielsen, MM, Tehler, D, Vang, S, Sudzina, F, Hedegaard, J, Nordentoft, I, Ørntoft, TF, Lund, AH, Pedersen, JS 2014, "Identification of expressed and conserved human noncoding RNAs", *RNA*, vol. 20, p. 236–251.

Niranjanakumari, S, Lasda, E, Brazas, R, Garcia-Blanco, MA 2002, "Reversible cross-linking combined with immunoprecipitation to study RNA–protein interactions in vivo", *Elsevier. Methods*, vol. 26 (2), pp. 182-190.

Ohto, MA, Floyd, SK, Fischer, RL, Goldberg, RB, Harada, JJ 2009, "Effects of APETALA2 on embryo, endosperm, and seed coat development determine seed size in *Arabidopsis*", *Sexual Plant Reproduction*, vol. 22(4), pp.277-289.

Orozco-Arroyo, G, Paolo, D, Ezquer, I, Colombo, L 2015, "Networks controlling seed size in *Arabidopsis*", *Plant Reprod.*, vol.28(1), pp.17-32.

Palovaara, J, Saiga, S, Wendrich, JR, Wout Hofland, NV, Schayck, JPV, Hater, F, Mutte, S, Sjollema, J, Boekschoten, M, Hooiveld, GJ, Weijers D 2018, "Transcriptome dynamics revealed by a gene expression atlas of the early *Arabidopsis* embryo", *Nat Plants*, vol.4(2):128.

Pauli, A, Valen, E, Lin, MF, Garber, M, Vastenhouw, NL, Levin, JZ, Fan, L, Sandelin, A, Rinn, JL, Regev, A, Schier, AF 2012, "Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis", *Genome Res.*, vol.22, pp.577–591.

Pinyopich, A, Ditta, GS, Savidge, B, Liljegren, SJ, Baumann, E, Wisman, E, Yanofsky, MF 2003, "Assessing the redundancy of MADS-box genes during carpel and ovule development", *Nature,* vol. 424(6944), pp.85-88.

Pignocchi, C, Minns, GE, Nesi, N, Koumproglou, R, Kitsios, G, Benning, C, Lloyd, CW, Doonan, JH and Hills, MJ 2009, “ENDOSPERM DEFECTIVE1 is a novel microtubule-associated protein essential for seed development in *Arabidopsis*”, *Plant Cell*, vol. 21 , pp. 90-105.

Portereiko, MF, Lloyd, A, Steffen, JG, Punwani, JA, Otsuga, D, Drews, GN 2006, “AGL80 is required for central cell and endosperm development in *Arabidopsis*”, *The Plant Cell*, vol. 18(8), pp. 1862-1872.

Popadin, K, Gutierrez-Arcelus, M, Dermitzakis, ET, Antonarakis, SE 2013, “Genetic and epigenetic regulation of human lincRNA gene expression”, *Am. J. Hum. Genet.*, vol. 93, pp. 1015–1026.

Ponting, CP, Oliver, PL, Reik, W 2009, “Evolution and functions of long noncoding RNAs”, *Cell*, vol. 136, pp. 629–641.

Prasad, K, Zhang, X, Tobón, E and Ambrose, BA 2010, “The *Arabidopsis* B-sister MADS-box protein, GORDITA, represses fruit growth and contributes to integument development”, *The Plant journal : for Cell and Molecular Biology*, vol. 62(2), pp. 203-214.

Quinn, JJ, Chang, HY 2015, “In situ dissection of RNA functional subunits by domain-specific chromatin isolation by RNA purification (dChIRP). In: Nakagawas S, Hiroses T (eds), “Nuclear Bodies and Noncoding RNAs”, *Methods in Molecular Biology*, vol. 1262, Humana Press, New York, NY.

Radoeva, T, Ten Hove, CA, Saiga, S, Weijers, D 2016, “Molecular characterization of *Arabidopsis* GAL4/UAS enhancer trap lines identifies novel cell-type-specific promoters”, *Plant Physiol*, vol. 171(2), pp. 1169-1181.

Riefler, M, Novak, O, Strnad, M, Schmülling, T 2006, “*Arabidopsis* cytokinin receptor mutants reveal functions in shoot growth, leaf senescence, seed size, germination, root development, and cytokinin metabolism”, *The Plant Cell*, vol. 18(1), pp.40-54.

Rinn, JL, Chang, HY 2012, " Genome regulation by long noncoding RNAs”, *Annu Rev Biochem.*, vol.81, pp.145-166.

Roxrud, I, Lid, SE, Fletcher, JC, Schmidt, ED and Opsahl-Sorteberg, HG 2007, “GASA4, one of the 14-member *Arabidopsis* GASA family of small polypeptides, regulates flowering and seed development”, *Plant and Cell Physiology*, vol. 48(3), pp. 471-483.

Sanda, SL and Amasino, RM 1996, “Ecotype-specific expression of a flowering mutant phenotype in *Arabidopsis thaliana*”, *Plant Physiol.*, vol. 111, pp. 641–644.

Salzman, J, Gawad, C, Wang, PL, Lacayo, N, Brown, PO 2012, “Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types”, *PLoS One*, vol.7, e30733.

Salzman, J, Chen, RE, Olsen, MN, Wang, PL, Brown, PO 2013, “Cell-type specific features of circular RNA expression”, *PLoS Genet.*, vol.9, e1003777.

Sado, T, Wang, Z, Sasaki, H, Li, E 2001, “Regulation of imprinted X-chromosome inactivation in mice by Tsix”, *Development*, vol.128, pp.1275–1286.

Schuettengruber, B, Cavalli, G 2009, “Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice”, *Development*, vol.136, pp.3531–3542.

Schruff, MC, Spielman, M, Tiwari, S, Adams, S, Fenby, N and Scott, RJ 2006, “The AUXIN RESPONSE FACTOR 2 gene of *Arabidopsis* links auxin signalling, cell division, and the size of seeds and other organs”, *Development*, vol. 133(2), pp.251-261.

Seo, JS, Sun, H-X, Park, BS, Huang, CH, Yeh, SD, Jung, C, Chua, NH 2017, “ELF18-INDUCED LONG-NONCODING RNA associates with Mediator to enhance expression of innate immune response genes in *Arabidopsis*”, *Plant Cell*, *vol.* 29, pp. 1024–1038.

Shin, JH and Chekanova, JA, 2014, “*Arabidopsis* RRP6L1 and RRP6L2 function in FLOWERING LOCUS C silencing via regulation of antisense RNA synthesis”, *PLoS Genet*, *vol.* 10(9):e1004612.

Shin, SY, and Shin, C 2016, “Regulatory noncoding RNAs in plants: potential gene resources for the improvement of agricultural traits”, *Plant Biotechnology Reports*, vol.10(2), pp.35-47.

Simon, JA, Kingston, RE 2009, “Mechanisms of polycomb gene silencing: knowns and unknowns”, N*at Rev Mol Cell Biol*, vol.10, pp.697–708.

Siomi, MC, Sato, K, Pezic, D and Aravin, AA 2011, “PIWI-interacting small RNAs: The vanguard of genome defence”, *Nat Rev Mol Cell Biol*, vol. 12, pp. 246–258.

Sorensen, MB, Mayer, U, Lukowitz, W, Robert, H, Chambrier, P, Jürgens, G, Somerville, C, Lepiniec, L and Berger, F 2002, “Cellularization in the

endosperm of *Arabidopsis thaliana* is coupled to mitosis and shares multiple components with cytokinesis", *Development*, vol. 129, pp. 5567-5576.

Somerville, C and Koornneef, M 2002, "A fortunate choice: the history of *Arabidopsis* as a model plant", *Nat Rev Genet*, vol. 3, pp. 883–889.

Sorenson, R and Bailey-Serres, J 2015, "Rapid immunoprecipitation of ribonucleoprotein complexes of plants", *Methods Mol Biol.*, vol. 1284, pp. 209-219.

Song, X, Sun, L, Luo, H, Ma, Q, Zhao, Y, Pei, D 2016, "Genome-Wide Identification and Characterization of Long Noncoding RNAs from Mulberry (*Morus notabilis*) RNA-seq Data", *Genes*, vol. 7(3), pp. 11.

Spitale, RC, Tsai, MC, Chang, HY 2011, "RNA templating the epigenome: long noncoding RNAs as molecular scaffolds", *Epigenetics*, vol. 6, pp.539–543.

Sreenivasulu, N and Wobus, U 2013, "Seed development programs: A systems biology-based comparison between dicots and monocots", *Annual Review of Plant Biology*, vol. 64, pp. 189-217.

Steffen, JG, Kang, IH, Portereiko, MF, Lloyd, A and Drews, GN 2008, "AGL61 interacts with AGL80 and is required for central cell development in *Arabidopsis*", *Plant Physiology*, vol. 148(1), pp. 259-268.

St Laurent, G, Wahlestedt, C, Kapranov, P 2015, "The Landscape of long noncoding RNA classification", *Trends Genet.*, vol. 31(5), pp. 239-251.

Sun, X, Shantharaj, D, Kang, X, Ni, M 2010, "Transcriptional and hormonal signaling control of *Arabidopsis* seed development", *Current Opinion in Plant Biology*, vol. 13(5), pp. 611-620.

Sunwoo, H, Dinger, ME, Wilusz, JE, Amaral, PP, Mattick, JS, Spector, DL 2009, "MEN *ε/β* nuclear-retained noncoding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles", *Genome Res*., vol.19, pp.347–359.

Swiezewski, S, Liu, F, Magusin, A, Dean, C 2009, "Cold-induced silencing by long antisense transcripts of an *Arabidopsis* Polycomb target", *Nature*, vol. 462, pp. 799–802.

Takahashi, N, Nakazawa, M, Shibata, K, Yokota, T, Ishikawa, A, Suzuki, K, Kawashima, M, Ichikawa, T, Shimada, H and Matsui, M 2005, "shk1-D, a dwarf *Arabidopsis* mutant caused by activation of the CYP72C1 gene,

has altered brassinosteroid levels", *The Plant Journal: for Cell and Molecular Biology*, vol. 42(1), pp. 13-22.

Terzi, CL and Simpson, GG 2009, "*Arabidopsis* RNA immunoprecipitation", *The Plant Journal*, vol. 59, pp. 163–168.

Thakur, N, Tiwari, VK, Thomassin, H, Pandey, RR, Kanduri, M, Göndör, A, Grange, T, Ohlsson, R, Kanduri, C 2004, "An antisense RNA regulates the bidirectional silencing property of the Kcnq1 imprinting control region", *Mol. Cell. Biol.*, vol.24, pp.7855–7862.

Tiwari, S, Schulz, R, Ikeda, Y, Dytham, L, Bravo, J, Mathers, L, Spielman, M, Guzmán, P, Oakey, RJ, Kinoshita, T and Scott, RJ 2008, "MATERNALLY EXPRESSED PAB C-TERMINAL, a novel imprinted gene in *Arabidopsis*, encodes the conserved C-terminal domain of polyadenylate binding proteins", *The Plant Cell*, vol. 20(9), pp.2387-2398.

Tian, Y, Lv, X, Xie, G, Zhang, J, Xu, Y, Chen, F 2018, " Seed-specific overexpression of AtFAX1 increases seed oil content in *Arabidopsis*", *Biochem Biophys Res Commun.*, vol. 500(2), pp.370-375.

Tsai, MC, Manor, O, Wan, Y, Mosammaparast, N, Wang, JK, Lan, F, Shi, Y, Segal, E, Chang, HY 2010, "Long noncoding RNA as modular scaffold of histone modification complexes", *Science*, vol. 329, pp. 689–693.

Venglat, P, Xiang, D, Wang, E and Dalta, R 2014, "Genomics of seed development: Challenges and opportunities for genetic improvement of seed traits in crop plants", *Biocatalysis and Agricultural Biotechnology*, vol. 3, pp. 24–30.

Villar, D, Flicek, P, Odom, DT 2014, "Evolution of transcription factor binding in metazoans – mechanisms and functional implications", *Nat. Rev. Genet.*, vol. 15, pp. 221–233.

Vinkenoog, R, Bushell, C, Spielman, M, Adams, S, Dickinson, GH and Scott, R 2003, "Genomic imprinting and endosperm development in flowering plants", *Mol Biotechnol.*, vol. 25(2), pp.149-84.

Wang, H, Wang, L, Erdjument-Bromage, H, Vidal, M, Tempst, P, Jones, RS, Zhang, Y, 2004, "Role of histone H2A ubiquitination in Polycomb silencing", *Nature,* vol. 431, pp.873–878.

Wang, A, Garcia, D, Zhang, H, Feng, K, Chaudhury, A, Berger, F, Peacock, WJ, Dennis, ES and Luo, M 2010, "The VQ motif protein IKU1 regulates

endosperm growth and seed size in *Arabidopsis*", *The Plant Journal : for Cell and Molecular Biology*, vol. 63(4), pp. 670-679.

Wang, KC, Yang, YW, Liu, B, Sanyal, A, Corces-Zimmerman, R, Chen, Y, et al. 2011, "A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression", *Nature*, vol. 472, pp. 120–124.

Wang, H, Niu, QW, Wu, HW, Liu, J, Ye, J, Yu, N, Chua, NH 2015, "Analysis of noncoding transcriptome in rice and maize uncovers roles of conserved lncRNAs associated with agriculture traits", *Plant J.*, vol.84, pp.404–416.

Wang, X, Ai, G, Zhang, C, Cui, L, Wang, J, Li, H, Zhang, J, Ye, Z 2016, "Expression and diversification analysis reveals transposable elements play important roles in the origin of Lycopersiconspecific lncRNAs in tomato", *New Phytologist*, vol. 209, pp. 1442–1455.

Wang, D, Qu, Z, Yang, L, Zhang, Q, Liu, ZH, Do, T, Adelson, DL, Wang, ZY, Searle, I, Zhu, JK 2017, "Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in Plants", *Plant J.*, vol. 90, pp. 133-146.

Wang, HV and Chekanova JA, 2017, "Long Noncoding RNAs in Plants", *Adv Exp Med Biol., vol.* 1008, pp. 133-154.

Washietl, S, Kellis, M, Garber, M 2014, "Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals", *Genome Res.*, vol. 24, pp. 616–628.

Wierzbicki, AT 2012, "The role of long noncoding RNA in transcriptional gene silencing", *Curr Opin Plant Biol*., vol. 15, pp. 517-522.

Weinhofer, I, Hehenberger, E, Roszak, P, Hennig, L and Kohler, C 2010, "H3K27me3 profiling of the endosperm implies exclusion of polycomb group protein targeting by DNA methylation", *PLoS Genet*., vol. 6, e1001152.

West, JA, Davis, CP, Sunwoo, H, Simon, MD, Sadreyev, RI, Wang, PI, Tolstorukov, MY, Kingston, RE 2014, "The long noncoding RNAs Neat1 and Malat1 bind active chromatin sites", *Mol Cell*, vol. 55, pp.791–802.

Wilusz, JE, Freier, SM, Spector, DL 2008, "3' End processing of a long nuclearretained noncoding RNA yields a tRNA-like cytoplasmic RNA", *Cell*, vol.135, pp.919–932.

Wu, H, Yin, QF, Luo, Z, Yao, RW, Zheng, CC, Zhang, J, Xiang, JF, Yang, L, Chen, LL 2016, “Unusual processing generates SPA lncRNAs that sequester multiple RNA binding proteins”, *Mol. Cell*, vol.64, pp.534–548.

Wu, H, Yang, L, Chen, LL 2017, “The diversity of long noncoding RNAs and their generation”, *Trends Genet*., vol.33(8), pp.540-552.

Xin, M, Wang, Y, Yao, Y, Song, N, Hu, Z, Qin, D, Xie, C, Ni, Z, Peng, H and Sun, Q, 2011, “Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing”, *BMC Plant Biol*, vol. 11, pp. 61.

Xing, YH, Yao, RW, Zhang, Y, Guo, CJ, Jiang, S, Xu, G, Dong, R, Yang, L, Chen, LL 2017, “SLERT regulates DDX21 rings associated with Pol I transcription”, *Cell* , vol.169, pp.664–678.e16.

Xiao, W, Brown, RC, Lemmon, BE, Harada, JJ, Goldberg, R and Fischer, RL 2006, “Regulation of seed size by hypomethylation of maternal and paternal genomes”, *Plant Physiology*, vol. 142(3), pp.1160-1168.

Xiao, J, Jin, R, Yu, X, Shen, M, Wagner, JD, Pai, A, Song, C, Zhuang, M, Klasfeld, S, He, C, Santos, AM, Helliwell, C, Pruneda-Paz, JL, Kay, SA, Lin, X, Cui, S, Garcia, MF, Clarenz, O, Goodrich, J, Zhang, X, Austin, RS, Bonasio, R, Wagner, D 2017, “Cis and trans determinants of epigenetic silencing by Polycomb repressive complex 2 in *Arabidopsis*”, *Nat Genet.*, vol.49(10), pp.1546-1552.

Yang, Y, Wen, L, Zhu, H 2015, “Unveiling the hidden function of long noncoding RNA by identifying its major partner-protein”, *Cell biosci*, Vol.5, pp.59.

Yin, QF, Yang, L, Zhang, Y, Xiang, JF, Wu, YW, Carmichael, GG, Chen, LL 2012, “Long noncoding RNAs with snoRNA ends”, *Mol. Cell*, vol.48, pp.219–230.

Young, RS, Marques, AC, Tibbit, C, Haerty, W, Bassett, AR, Liu, JL, Ponting, CP 2012, “Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome”, *Genome Biol. Evol.*, vol. 4, pp. 427–442.

Yoon, JH, Abdelmohsen, K and Gorospe, M 2013, “Posttranscriptional gene regulation by long noncoding RNA”, *J. Mol. Biol.,* vol. 425, pp. 3723–3730.

Yoshida, N, Yanai, Y, Chen, L, Kato, Y, Hiratsuka, J, Miwa, T, Sung, ZR, Takahashi, S 2001, “EMBRYONIC FLOWER2, a novel Polycomb group protein homolog, mediates shoot development and flowering in *Arabidopsis*”, *Plant Cell*, vol.13, pp.2471–2481.

Zhang, Y, Liang, W, Shi, J, Xu, J, Zhang, D 2013a, “MYB56 encoding a R2R3 MYB transcription factor regulates seed size in *Arabidopsis thaliana*”, *J Integr Plant Biol.*, vol. 55(11), pp.1166-78.

Zhang, Y, Zhang, XO, Chen, T, Xiang, JF, Yin, QF, Xing, YH, Zhu, S, Yang, L, Chen, LL 2013b, “Circular intronic long noncoding RNAs”, *Mol. Cell*, vol.51, pp.792–806.

Zhang, YC, Liao, JY, Li, ZY, Yu, Y, Zhang, J, Li, Q, Qu, L, Shu, W, Chen, Y 2014a, “Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice”, *Genome biology*, vol. 15(12), pp. 512.

Zhang, XO, Yin, QF, Wang, HB, Zhang, Y, Chen, T, Zheng, P, Lu, X, Chen, LL, Yang, L 2014b, “Species-specific alternative splicing leads to unique expression of sno-lncRNAs”, *BMC Genomics*, vol.15, pp.287.

Zhang, XO, Wang, HB, Zhang, Y, Lu, X, Chen, LL, Yang, L 2014c, “Complementary sequence-mediated exon circularization”, *Cell*, vol.159, pp.134–147.

Zhou, Y, Zhang, X, Kang, X, Zhao, X, Zhang, X and Ni, M 2009, “SHORT HYPOCOTYL UNDER BLUE1 associates with MINISEED3 and HAIKU2 promoters in vivo to regulate *Arabidopsis* seed development”, *The Plant Cell*, vol. 21(1), pp.106-117.

Zhou, Z, Wang, Z, Li, W, Fang, C, Shen, Y, Li, C, Wu, Y, Tian, Z 2013, “Comprehensive analyzes of microRNA gene evolution in paleopolyploid soybean genome”, *Plant Journal*, vol.76, pp. 332–344.

Appendices

8.1 Supporting documents

8.1.1 Chapter 1: Introduction

Table S1. Genes demonstrated to have a role in seed development in *Arabidopsis thaliana*

Gene name	Abbreviation	Protein Function	Phenotype	Reference
EARLY FLOWERING IN SHORT DAYS	*EFS*	Contributor for H3K36 methylation	Larger embryo in mutants	Cheng et al., 2018
ARABIDOPSIS FATTY ACID EXPORT 1	*AtFAX1*	Mediate the fatty acid export from plastid.	Larger seeds in over-expression plant	Titan et al., 2018
MYB56	*MYB56*	R2R3 MYB transcription factor	Larger seeds in over-expression plant	Zhang et al., 2013a
APETALA2	*AP2*	AP2 domain transcription factor	Larger seeds in mutants	Ohto et al., 2009
FERTILIZATION INDEPENENT SEED 2	*FIS2*	Polycomb group protein	Reduced embryo development	Sun et al., 2010
FERTILIZATION-INDEPENDENT ENDOSPERM	*FIE or FIS3*	Polycomb group protein	Reduced embryo development	Sun et al., 2010
MEDEA	*MEA or FIS1*	Polycomb group protein	Reduced embryo development	Sun et al., 2010
MULTICOPY SUPPRESSOR OF IRA	*MSI1*	Polycomb group protein	Reduced embryo development	Sun et al., 2010
METHYL TRANSFERASE 1	*MET1*	DNA methyl transferase	Larger seeds (mutant maternal plant) or smaller seeds (mutant paternal plant)	Sun et al., 2010
HAIKU1	*IKU1*	VQ motif protein	Smaller seeds in mutants	Wang et al., 2010

GORDITA or AGAMOUS-LIKE 63	*GORDITA or AGL63*	Bsister MADS-box transcription factor	Larger fruits in mutants	Prasad et al., 2010
AUXIN BINDING PROTEIN 1	*ABP1*	Auxin binding protein	Abnormal embryo morphology	Sun et al., 2010
ARABIDOPSIS THALIANA CULLIN 1	*AtCULLIN1*	Protein binding	Abnormal embryo morphology	Sun et al., 2010
ARABIDOPSIS HISTIDINE PHOSPHOTRANSFER PROTEIN	*AHP*	Cytokinin single transducer	Larger seeds in *ahp 2,3,5* triple mutants	Sun et al., 2010
CYTOKININ INDEPENDENT 1	*CKI1*	Histidine kinase without cytokinin perception domain	Larger seeds in one of mutant alleles	Sun et al., 2010
SHORT HYPOCOTYL UNDER BLUE 1	*SHB1*	Transcription co-activator	Larger seeds in over-expression plants	Zhou et al., 2009
CYP78A9		P450 monooxygenase family protein	Smaller seeds in mutants	Adamski et al., 2009
KLUH or CYP78A5	*KLU*	P450 monooxygenase family protein	Smaller seeds in mutants	Adamski et al., 2009
MATERNALLY EXPRESSED PAB C-TERMINAL	*MPC*	poly(A) binding protein	Smaller seeds in *MPC* RNAi knockdown plants	Tiwari et al., 2008
DIANA or AGAMOUS-LIKE61	*AGL61*	MADS-box transcription factor	No seed in mutants	Steffen et al., 2008
AGAMOUS-LIKE62	*AGL62*	MADS-box transcription factor	No seed in mutants	Kang et al., 2008
RETARDED GROWTH OF EMBRYO 1	*RGE1*	bHLH transcription factor	Smaller and shrivelled seeds in mutants	Kondou et al., 2008
DA1	*DA1*	ubiquitin receptor	Larger seeds in over-expression planta	Li et al., 2008
GIBBERELLIC ACID-STIMULATED ARABIDOPSIS 4	*GASA4*	gibberellin-responsive protein	Larger seeds in over-expression plant	Roxrud et al., 2007
FEM111 or AGAMOUS-LIKE80	*AGL80*	MADS-box transcription factor	No seed in mutants	Portereiko et al., 2006

DECREASE IN DNA METHYLATION 1	*DDM1*	Chromatin remodelling factor	Larger seeds (mutant maternal plant) or smaller seeds (mutant paternal plant)	Xiao et al., 2006
MEGA INTEGUMENTA or AUXIN RESPONSE FACTOR 2	*MNT or ARF2*	Auxin-responsive element binding transcription factor	Larger seeds in mutants	Schruff et al., 2006
ARABIDOPSIS HISTIDINE KINASE	*AHK*	Cytokinin receptor	Larger seeds in *ahk2,3,4* triple mutants	Riefler et al., 2006
BETA-XYLOSIDASE 3	*BX3*	beta-xylosidase	Smaller seeds in mutants	Minic et al., 2006
SUCROSE-PROTON SYMPORTER 5	*SUC5*	sucrose transporter	Decrease the dry weight of seed in mutants	Baud et al., 2005
HAIKU2	*IKU2*	Leucine-rich repeat receptor kinase	Smaller seeds in mutants	Luo et al., 2005
MINISEED3	*MINI3*	WRKY transcription factor	Smaller seeds in mutants	Luo et al., 2005
TRANSPARENT TESTA GLABRA 2	*TTG2*	WRKY transcription factor	Smaller seeds in mutants	Garcia et al., 2005
SHRINK 1 or CYP72C1	*SHK1*	P450 monooxygenase family protein	Smaller seeds in over-expression plants	Takahashi et al., 2005
SEEDSTICK or AGAMOUS-LIKE11	*STK*	MADS-box transcription factor	Small seed in mutants	Pinyopich et al., 2003
EXTRA SPOROGENOUS CELLS or ECXESS MICROSPOROCYTES 1	*EXE or EMS1*	Leucine-rich repeat receptor kinase	Smaller seeds in mutants	Canales et al., 2002
DEMETER	*DME*	DNA 5-methyl cytosine demetthylase	Nonviable seeds in mutants	Choi et al., 2002
TRANSPARENT TESTA 16 or ARABIDOPSIS BSISTER	*TT16 or ABS*	B(S) MADS-box transcription factor	Larger seeds in mutants	Nesi et al., 2002
AINTEGUMENTA	*ANT*	AP2-like transcription factor	No seed in mutants	Mizukami and Fischer 2000
mutants: loss of gene function				

8.1.2 Chapter 5: Identification of PRC2-associated Long noncoding RNA in *Arabidopsis thaliana* Siliques

Table S1. Primers used in this chapter

Primers	Sequences (5'-3')
linc_23526_F	TTT GAA GGT GCT AGA CGG GT
linc_23526_R	TCG ACA CCA TCC ACA TCC AT
linc_23618_F	TTA TAT GAC AGG GCC GCT CA
linc_23618_R	GGC CAT AAT GTT TCC CCT TGA
linc_28194_F	GAA TCG CTT CCT CAC ATA GCT
linc_28194_R	ACA TAA GAA AAC CAA GGC CGT
linc_29066_F	TGA AAG CAG GCA GTC AAA GG
linc_29066_R	CCC AGG TTC GAA ACA CAC AC
linc_34938_F	ACT TAT GTC GGT CGC TTT GTG
linc_34938_R	CCA ACC AAG CTC CAT CAA CC
linc_11274_F	GGA TCC ATG AGC AAG TAT CAC A
linc_11274_R	ACC AGT AAG ATT CTC CAC TAG CT
linc_11427_F	AAT AGA GAG CGG CCA AAA CG
linc_11427_R	GCT TAT GTG TGGT GGT GTG G
SALK_047543_LP	GGT CCA ATG AAC ATC GTT GAC
SALK_047543_RP	CAT GTT TTG TTC TTA AAA TAC ATG C
SALK_038231_LP	TGA AGG GAC AAG AGG TTC AAG
SALK_038231_RP	TGT CAA CAG TTT CAA CAT GAC AAC
SALK_095819_LP	TCA ATT TGT GAC TTA TGT CTA TCA TTG
SALK_095819_RP	TGA GTT GTG GAC CCT TTG TTG
SALK_102768C_LP	GAA GGT TAA ATA ACC GCA TTA TTG
SALK_102768C_RP	GGT TGA CTG GAA CTG ATT TCG
SALK_058251_LP	CCA CTG TTG AAT GTT ATG CAG G
SALK_058251_RP	CAG GAT TTA TAT GCT AAC AGA GTT AAG C
FLAG_205A06_LP	TGG GTG AGT TAA AAG CAT TCG
FLAG_205A06_RP	ATG TGG CGT AGT TTT ACT GGG
FLAG_395F03_LP	AGC TCA TAC CCA TGA ATC TCG
FLAG_395F03_RP	TCA TCG AAT GGA AAA ACG AAC
FLAG_497A02_LP	CAT TGG TCT CGA GCT TCT CTG
FLAG_497A02_RP	GAT GGC ACA CTG TTT CCT TTG
FLAG_269H08_LP	CCC TCT TGG TGA AGT AGA GGG
FLAG_269H08_RP	TTC ATC ATA TTC ACT GGA TTG ATT G
AT2G25450_LP	CGG CTC TTC ACC TCC ATT TG
AT2G25450_RP	TCC TCC TAA TCC CGA AGC AC
AT2G25700_LP	CGT AAG GTC AAG AAG TCC ACT G
AT2G25700_RP	AGT ACT GCA AGA AAC ACG TTG A
AT4G29640_LP	TCT TCC AAG ACC CGT GCT AC

AT4G29640_RP	GAC AGT TAT CGC GCT CCA TG
*AT1G73610*_LP	ACG GTA GAG ACT TTA TAG GTG GA
*AT1G73610*_RP	GCG AAA CAA ACA CCA GTC GT
AT1G76500_LP	GTT ACC ACT TAC GCT CGC AG
AT1G76500_RP	CCT TGA ACG CCG GAA AGA AA
AT3G59010_LP	CTG ATC CGA CCC GTA ATG CT
AT3G59010_RP	CAG AAG TAA CAG AGG CGG CT
AT1G03445_LP	TCT CCA AGC TGT GTT GTC CA
AT1G03445_RP	CCA ATC AAA GTC TTC GGC GA
AT1G75900_LP	ACC AGC ATA CCT AGA TCC GA
AT1G75900_RP	CTG CCT GCG ACC AAT AAG AA
AT1G76290_LP	CAC GAT TTC TCC TCC GCA TG
AT1G76290_RP	GAC AAG GAA GCC ACT GAA GC
lncRNA_13468_LP	AGC CTT TCT CTT TCT TCT TCC T
lncRNA_13468_RP	GAG AGA ACA TGT GGG TGA ACA
lncRNA_17943_LP	TGC TCT TAG AGT TAT TGT GG
lncRNA_17943_RP	CTT AAC AGA TTT ACA CGT CTC
lncRNA_17992_LP	AGC GAG AAG GAT CAG TTG GA
lncRNA_17992_RP	AGA TCC AGT GAA GAG GTC CG
lncRNA_20351_LP	TCC TGT CAT GCA AGA AAC CT
lncRNA_20351_RP	TGA CCA CCA TTG ACT CAC TCA
lncRNA_32863_LP	TGT TGT CAT GTT GAA ACT GTT GA
lncRNA_32863_RP	ACC ATC GAA GTA ACT CAC ACA T
lncRNA_19163_LP	TGA GCC TCT TCC TTC ACC AT
lncRNA_19163_RP	CAG ATG AAG AAG ACG ACG GTA TC
lncRNA_528_LP	TAG TTT TAT CGA CCG GAC CG
lncRNA_528_RP	CAA CAA AGG GTC CAC AAC TCA
lncRNA_12840_LP	AAC CCA AAG TGA GCC CTC TT
lncRNA_12840_RP	TCT TCT TTT ATG GCA GTT GGT CT
lncRNA_13344_LP	GGG CCA TTT AGT TGT CAG TCA
lncRNA_13344_RP	TAG CAC TAC CAT TCC ACG GA
lncRNA_13393_LP	AAC GCT CCA AAC AAT TAA TAT GG
lncRNA_13393_RP	ATA AGA TTT GTT TAG TTG AC

8.1.3 Chapter 6: Maternal Control of Seed Size by a Long noncoding RNA in *Arabidopsis thaliana*.

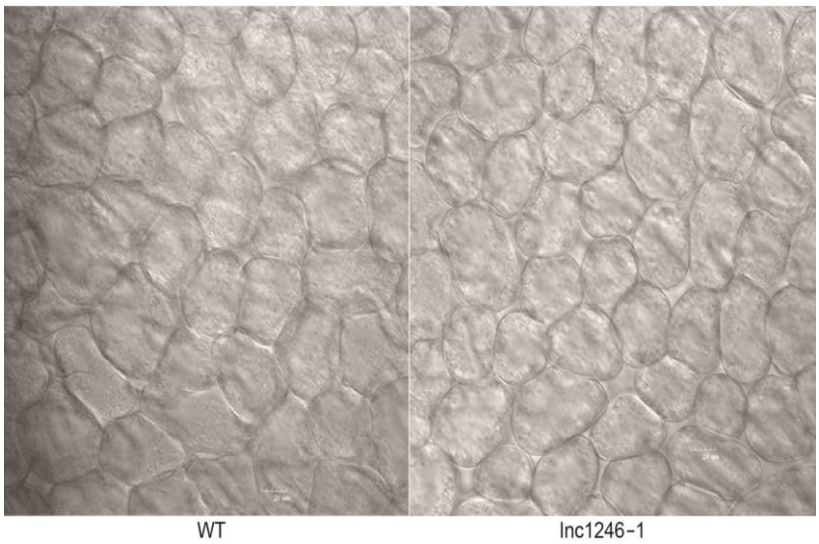
8.1.3.1 Table S1. Primers used in this chapter

Primers	Sequences (5'-3')
LINC.TCONS_2215_F	GGA CGA GAA TTT GAC TCC ACG
LINC.TCONS_2215_R	ACC CTC TTT CTT GTT TCG TCG

exonAS.TCONS_244_F	CCT TCA AGA TCT CTC CCG TC
exonAS.TCONS_244_R	TCA GAT CAC CCG ACA CTC TC
exonAS.TCONS_1177_F	GTG CTT TCT TGA GGG CTA CG
exonAS.TCONS_1177_R	CGA GGC CAT GAT CGC GGA AG
LINC.TCONS_719_F	CAG TAA AGC CCA TTG ACA AGG
LINC.TCONS_719_R	CGA TTG AGA GAG GGA CCG TG
intronAS.TCONS_120_F	TGC ACC TGA CAC TAT TCT GC
intronAS.TCONS_120_R	TTG GCG ATT TCC TGA GTT GC
intronAS.TCONS_682_F	TGG TGT TCG GAT GGT GTA TTG
intronAS.TCONS_682_R	CTA GGG TGA ATG CAT AGG GAC
intronAS.TCONS_976_F	CAA CAA CCA ACC AAA CCA CC
intronAS.TCONS_976_R	TGC AGC CTA ACC ATC TGT GAG
intronAS.TCONS_1171_F	AGC CTC AAT TCA CGG GTT AAC
intronAS.TCONS_1171_R	ACA GAA GCA AGG TCC CTC AG
intronAS.TCONS_2182_F	CCA TGG CCT CTT CAA CCA AG
intronAS.TCONS_2182_R	GTG TTG TGT CGA TCG TGC G
intronAS.TCONS_2762_F	TGG TTT GAG AAA GGA GCA CC
intronAS.TCONS_2762_R	CTT AGG TTA GGA GGG CAT TGC
SALK_207384_F	TTG AGG ACC AAG ATC CAC ATC
SALK_207384_R	TCT GCT CGG CTT TAT TTT CAC
SALK_LB1.3	ATT TTG CCG ATT TCG GAA C
exonAS.TCONS_1246_F	GTG CTG TGC TCC ATG AAA GG
exonAS.TCONS_1246_R	GCT GCT CGT GTA GTT CTT GA
exonAS.TCONS_1246_RT	TTG GGT CGT GTC AAG GTT TG
Sand (AT2G28390)_F	CAG ACA AGG CGA TGG CGA TA
Sand (AT2G28390)_R	GCT TTC TCT CAA GGG TTT CTG GGT
Actin1_F	GTC TCG AGA GAT GAC TCAG ATC ATG TTT GAG
Actin1_R	GGC GCG CCA CAA TTT CCC GTT CTG CGG TAG
Pdf2 - F	TCC ACA GCT TTC TCC CTC AC
Pdf2 - R	CGG CTT TCT ATC ATT GCT CGT
At3g12940_F	TCT GCA ATC TCC TGA ACT CGT
At3g12940_R	TCA TCG TCC CTC AAT CCC AG
Kluh - F	GGT ACG GCA GTT TTG GGA TG
Kluh - R	TGA TGT CTT GCT TGG CTT GC
mini3-F	ATC GCT GCA TTG TCT TCA CC
mini3-R	TCG TTG CAA TCT CTC CAG GA
iku1-F	GCC ACA GTC TCA TCC TCA GT
iku1-R	TCA TGA CCT GGC TGC ATT TG
iku2-F	GCT GCT AAA GGG CTG GAG TA
iku2-R	GCT TCT TCC CTG TCA CCA AC
ttg2-F	ATT CCG GTT GCA AGA GTA GC
ttg2-R	ATA CGC ATT GCC TCC TAC CA
lincRNA9137_F	GAT CGT ATG ATC CCC CGG ATT C
lincRNA9137_R	TAA ATG CAG ATC CCG GTG TAG G

linc_29066_F	TGA AAG CAG GCA GTC AAA GG
linc_29066_R	CCC AGG TTC GAA ACA CAC AC
linc_34938_F	ACT TAT GTC GGT CGC TTT GTG
linc_34938_R	CCA ACC AAG CTC CAT CAA CC

8.3.1.2 Figure S6. The cell size in the sub-epidermal cell layers of wild type and *LNCRNA_1246* mutant



8.2 Data Repository

The empirical findings reported in this thesis are publicly available. This data is stored, maintained and updated if necessary, in my shared Dropbox™ folder:

https://www.dropbox.com/sh/ecfhx0osb3rh5up/AABhunEJ4_UwglRaC8jaCS7Da?dl=0

To contact me, please send an email to the following address:
trungcnsinh@gmail.com

Regards,

Trung Do.